



Review

Advances in Artificial Intelligence (AI) Models and Generative Algorithms Represent a New Paradigm for Genomics Research

Du Hyeong Lee ^{1,2} , Eun Gyung Park ^{1,2} , Yun Ju Lee ^{1,2} , Hyeon-su Jeong ^{1,2} , Hyun-Young Roh ^{1,2} ,
Ga-ram Jeong ^{1,2} , Sang-Woo Kim ^{2,3} and Heui-Soo Kim ^{2,3,*}

¹ Department of Integrated Biological Science, Pusan National University, Busan 46241, Republic of Korea; doo2080@naver.com (D.H.L.); ehdtodt@pusan.ac.kr (E.G.P.); lsg5821@naver.com (Y.J.L.); tbd97@pusan.ac.kr (H.-s.J.); susan9416@naver.com (H.-Y.R.); gracefka@naver.com (G.-r.J.)

² Institute of Systems Biology, Pusan National University, Busan 46241, Republic of Korea; kimsww@pusan.ac.kr

³ Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 46241, Republic of Korea

* Correspondence: khs307@pusan.ac.kr

Abstract

Genomics has developed in step with progress in computing. As computational capabilities have grown, analyses have expanded from simple statistics to artificial intelligence (AI)-based approaches within genomics. The decline in sequencing costs has led to the accumulation of diverse genomic datasets, rapidly accelerating AI for genomic analysis. AI models are now developed and applied across many functional domains, including the prediction of transcription factor binding sites, epigenetic elements, DNA methylation, and noncoding sequence functional annotation. With the maturation of architectures such as deep neural networks, convolutional neural networks, recurrent neural networks, and transformers, many genomic models now accommodate longer inputs, capture long-range context, and integrate complex multi-omics data, thereby steadily improving predictive accuracy. Moreover, the emergence of generative AI has enabled models that can simulate and design genomic sequences. The introduction of generative AI into genomics goes beyond inferring function to the capability of replicating functional genomes. These advances will help advance genome interpretation and accelerate our ability to chart and navigate the genomic landscape.

Keywords: machine learning; artificial intelligence; generative algorithms; genomics; deep learning; bioinformatics



Academic Editors: Fei Guo and Paolo Uva

Received: 5 September 2025

Revised: 6 November 2025

Accepted: 10 November 2025

Published: 11 November 2025

Citation: Lee, D.H.; Park, E.G.; Lee, Y.J.; Jeong, H.-s.; Roh, H.-Y.; Jeong, G.-r.; Kim, S.-W.; Kim, H.-S. Advances in Artificial Intelligence (AI) Models and Generative Algorithms Represent a New Paradigm for Genomics Research. *Int. J. Mol. Sci.* **2025**, *26*, 10925. <https://doi.org/10.3390/ijms262210925>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid development of artificial intelligence (AI) is transforming life for people in a variety of different fields. AI models quickly augment traditional search engines, tools for data exploration, drafting, and administrative assistance [1–3]. This paradigm involves the replacement of traditional statistical analysis tools with faster and more efficient mechanisms not only in daily life but also in many areas of science, especially biology. AI is being used in a variety of fields of biology, such as determining the prognosis of cancer patients, developing new diagnostic tools, and predicting the outbreak of infectious diseases [4–6]. Current AI is performing roles ranging from simple data classification to generalizing patterns and extracting key information from large datasets generated by high-throughput molecular technologies [7]. Traditionally, genome analysis has included some types of tasks such as classifying short sequencing reads, using GWAS to identify polymorphisms

associated with phenotypic traits and diseases, and applying eQTL analysis to pinpoint genomic loci that influence gene-expression levels [8–10]. However, advances in computing and machine learning have spurred a growing body of work that uses deep-learning algorithms, such as artificial neural networks (ANNs), to infer phenotype-associated genes and genomic variants, predict protein functions, and model the structure of the genome [11]. In particular, with the development of the transformer architecture and self-attention-based systems exemplified by AlphaFold in analyzing protein structures, AI has been used to predict protein structures using large-scale data and for whole-genome research since the 2020s, and the use of AI in genome research is increasing [12,13]. Since then, AI models have been employed in the field of pharmaceutical research, including the development of treatments and vaccines for the novel coronavirus (SARS-CoV-2), and the use of AI in biology has seen a further increase [14]. The 2024 Nobel Prize in Chemistry was awarded to David Baker, Demis Hassabis, and John M. Jumper for developing AlphaFold 2, thus further raising awareness of AI in the context of biological research [15]. AI models now classify genomic data to infer disease risk and predict structure; they also synthesize novel gene or genome sequences conditioned on user prompts [16]. In this review, we will examine the advancement of genomics research, the use of AI, and genomics research using generative AI (Figure 1).

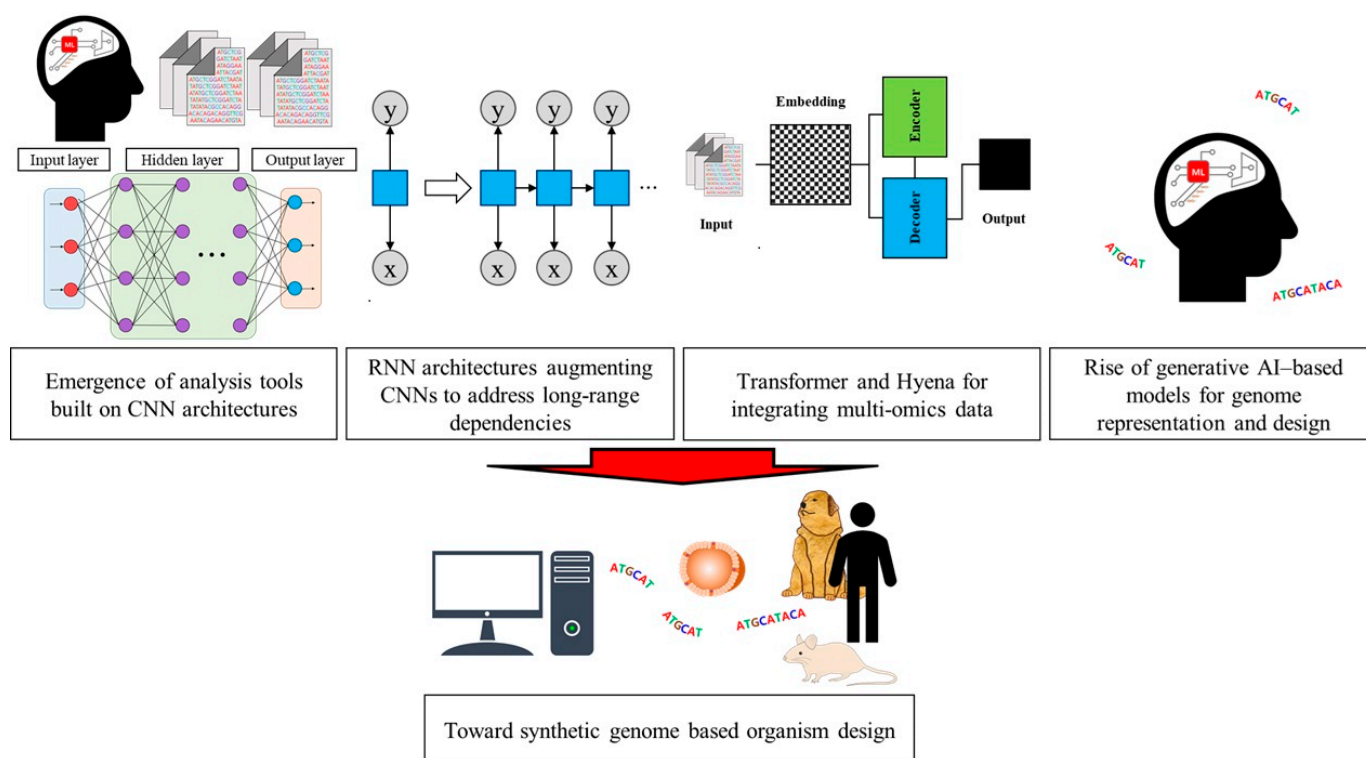


Figure 1. Schematic illustration of the development of genomics and the advancement of AI-based genomics research.

2. Machine Learning and Deep Learning Algorithms Used in Genomics Research

With developments in next-generation sequencing technologies driving down sequencing costs, genomic data have accumulated across diverse domains. Vast resources have been assembled, including consensus sequences from multiple organisms, genomes of newly characterized species, ChIP-seq/ATAC-seq datasets that capture interactions among regulatory factors, and CLIP-seq profiling protein–RNA binding [17,18]. As these datasets have grown, GWAS have been widely conducted to explore relationships between

genetic and genomic variation and phenotype and to analyze genomic function [9]. In parallel, advances in bioinformatics tools have made computing-centric analyses, from detecting SNPs associated with disease to predicting alternative DNA conformations in silico [19,20]. During this period, early machine learning algorithms emerged. These approaches, grounded in statistical modeling to maximize likelihood or in similarity-based rules for binary classification, include logistic regression, random forests (RF), boosting methods, k-nearest neighbors (k-NN), support vector machines (SVM), and Naïve Bayes (NB) [21–26]. Such models have been applied in genomics to infer SNP function and elucidate genotype–phenotype links, supporting applications such as cultivar development and the discovery of disease-causing variants (Table 1) [27–29].

Table 1. Early machine learning algorithms and application of genomic research.

Algorithm	Method	Application	References
Logistic regression	estimates the probability	Detection of SNP-SNP/gene interaction; Breeding and selection	[29,30]
Random Forest	Generate random classification and regression trees	Disease-associated SNP detection (Cancer, Alzheimer’s disease)	[31,32]
k-nearest neighbor	Utilize the k closest training instances in the feature space	SNP-SNP interaction detecting Microarray data analysis	[33–35]
Boosting machine	regression and classification that iteratively reduce residual	SNP pattern analysis for disease prediction Disease diagnosis (Cancer, Alzheimer’s disease)	[23,36] [37–39]
Naïve bayse	Bayes’ theorem to relate prior and posterior probabilities	Select biomarker SNP analysis of disease diagnosis (Alzheimer’s disease)	[27,40]
Supporting vector machine	Hyperplane that maximizes the margin between classes	Disease diagnosis Predict cancer-associated genes	[28,41,42]

With developments in computing, ANNs began to be applied in earnest. ANNs are models that solve problems by adjusting connection weights among multilayer nodes, an abstraction inspired by biological neural networks [11,43]. Although early performance did not clearly surpass techniques such as SVMs or NB due to limited computational resources, improvements in hardware and the accumulation of data led to the rapid spread of deep neural architectures such as deep neural network (DNN)/multi-layer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN)/long short-term memory (LSTM) in genomics after 2015 [44–46]. Furthermore, the advent of the transformer architecture catalyzed large-scale natural language models, and self-attention-based systems exemplified by AlphaFold, ushering in a new paradigm after 2020 in which AI research increasingly integrates diverse, large-scale datasets [12,47]. As a result, the accumulation of massive genomic data and dramatic gains in computational power accelerated the adoption of AI in genomics and drove substantial innovations in genomic function analysis.

2.1. Learning Paradigms and Characteristics of Deep Learning Models Used in Genomic Research

Recently, deep learning-based models used in genomics are categorized by learning paradigm into supervised, unsupervised, and semi-supervised learning. In supervised

learning, genomic data are accompanied by labels or annotations such as transcription start sites, transcription termination sites and splice sites. In contrast, unsupervised learning discovers latent patterns from large, unlabeled datasets [48–50]. In genomics, supervised learning trains predictors using known biological annotations, while unsupervised learning can be applied to uncover the structure of extensive variant and sequence data [51]. Supervised learning underlies many deep learning-based early genomic analysis models such as DeepBind. It often achieves strong predictive performance and is well-suited to fine-tuning of pretrained models. Still, it can be limited by the difficulty of data collection and the risk of overfitting. By contrast, unsupervised learning is easier to scale in terms of data acquisition and provides a foundation for pretrained models such as DNABERT, yet it carries the risk of learning spurious patterns [52–54]. Across these learning processes, the most critical steps are data acquisition and preprocessing, such as normalization and length handling. As dataset size increases, the risks of overfitting and underfitting generally diminish, and appropriate preprocessing can substantially reduce computational cost [54]. Selecting an architecture that matches the analysis objective and input data characteristics is essential, as each deep learning model entails its own advantages and disadvantages.

2.2. Characteristics of Major Deep Learning Architectures Used in Genomic Research

DNN and MLP apply successive nonlinear transformations to vector inputs through multiple hidden layers. They are fast and relatively easy to implement and modify. They can ingest heterogeneous input types and are straightforward to integrate with other deep-learning architectures [44,55]. Using at least 2–3 hidden layers is recommended, and employing more than 100 hidden layers can be effective. Compared with traditional machine-learning models such as NB, k-NN, RF, and SVM, DNNs often achieve superior accuracy. However, because all neurons are fully connected, the number of parameters grows combinatorially with input dimensionality, as in high-dimensional genomic data, so achieving reasonable performance can require substantial computation time [56,57]. Consequently, in genomics, standalone DNNs are more often used in auxiliary roles rather than ingesting large raw sequences directly, for example, by taking functional/disease-association features such as GO, PPI, PathDIP, KEGG as inputs to predict aging-related genes, or by learning short (<100 bp) sequence windows using enhancer-related histone-modification signals as in EP-DNN [58,59]. In short, while DNNs are simple to build and can outperform conventional statistical models in accuracy, they offer limited parameter efficiency and, due to their fully connected design, are not resource efficient.

CNNs replace the fully connected structure of DNNs with convolutional layers that exploit local patterns and weight sharing, yielding high parameter efficiency and fast training relative to model size [60,61]. However, CNNs typically require fixed-length inputs, which can introduce information loss and, for large contexts, may entail longer training times [62]. In genomics, CNNs have been used for tasks such as predicting the binding specificities of DNA/RNA-binding proteins in DeepBind and DeeperBind, and annotating functions of noncoding DNA regions in Basset and DanQ [53,63–65]. As deep learning methods have been applied to genomics, many supervised models have been built on CNNs. Owing to their relatively low implementation complexity and strong accuracy, CNN-based approaches remain widely used. That said, to address long training times in long sequences and limitations in modeling long-range dependencies, CNNs are now often combined with other architectures.

RNNs and their variant LSTM can model long-range dependencies more precisely than CNNs, and have thus been used to predict interactions between distantly spaced nucleotides. Their strength with variable-length inputs makes them well-suited to genomic data [66]. RNNs have been employed in model DeepZ to predict Z-DNA structure,

and LSTMs have been employed in models such as AttentiveChrome to predict chromatin interactions [67,68]. Because RNNs handle larger inputs and long-range dependencies more effectively than CNNs, many CNN-based models have been augmented with recurrent components.

Transformer architectures learn long-range interactions and global context effectively via self-attention and, as with BERT, a natural language processing model, are well suited to parallelization and to pretraining and transfer learning. Because they integrate heterogeneous data readily, many AI models developed since 2020 have adopted transformer architectures [69,70]. In genomics, supervised models such as Enformer target enhancer-associated prediction, while pretrained approaches like DNABERT, using k-mer tokenization, are widely used. These models enable the integration of diverse omics signals, including gene-expression regulation, transcription-factor binding sites, and chromatin accessibility [54]. With their unique scalability and appropriate fine-tuning, DNABERT can infer a variety of biological factors, including chromatin marks, transcription factor binding domains, and genome functions. However, transformers typically require substantially more training data and computational resources than other models [71]. Consequently, many studies extensively fine-tune pretrained models such as DNABERT to obtain task-optimized analyzers.

Hyena is an architecture designed to relax the transformer's relatively short effective context length, efficiently handling ultra-long-range context while maintaining single-nucleotide resolution. It has the potential to achieve high performance with comparatively few parameters, but it can be challenging to fine-tune [72]. Hyena has been applied to the genome-generative model EVO, where it demonstrated strong performance [16]. Although Hyena can accept very large inputs and thereby capture genomic characteristics well, it still requires extensive validation. Models currently used for genome analysis are being developed using various deep learning algorithms. The development trend is evolving toward building complex models by integrating large amounts of data and omics data over time (Table 2).

Table 2. Methods and features of deep learning architectures used in genomic research.

Model Architecture	Method	Strengths	AI Model of Applied Genomics
DNN/MLP	Fully connected layer structure	Simple and fast to implement Applicable to small datasets	EP-DNN
CNN	Convolutional, pooling, and fully connected layer structure	Learns local sequence patterns Efficient resource utilization through weight sharing Fast training	DeepBind DeeperBind Basset DanQ
RNN/LSTM	Recurrent layer structure	Strong in position dependence Variable-length input possible	AttentiveChrome SG-LSTM-FRAME

Table 2. Cont.

Model Architecture	Method	Strengths	AI Model of Applied Genomics
Transformer	Self-attention mechanism	Capturing long-range interactions; Parallel processing Easily extended to pretraining	Enformer DNABERT Nucleotide Transformer
Hyena	long-context convolutional sequence model	Handles long sequences Maintains single-base resolution High performance with fewer parameters and memory efficient	HyenaDNA EVO

3. Deep Learning Models Being Developed and Utilized in Various Fields of Genomic Research

Since deep learning demonstrated its potential in biology, a wide array of models has been developed and applied in genomics, centered on predicting the functional consequences of sequence variation. In particular, AI is actively developed and leveraged in areas that are difficult to resolve solely by experimentation or by existing annotated reference resources, including sequence-based prediction of transcription factor binding sites, regulation of gene expression (promoters, enhancers), epigenetic marks, alternative splicing sites, functions of noncoding RNAs, and detection of alternative DNA conformations.

3.1. Prediction of Binding Regions Between Nucleotides and Proteins

Predicting the binding regions of proteins such as transcription factors or regulatory factors are central to understanding gene expression, translation, and alternative splicing in genomics. DeepBind was the first to apply a deep learning approach in this prediction, training CNNs on data such as ChIP-seq and CLIP-seq to estimate binding propensity. And Basset used RNA-seq data to learn accessibility-based features to predict transcription factor binding potential in noncoding regions [53,64]. Likewise, the CNN-based BPNet inferred transcription factor binding motifs and captured them at base-level resolution, improving accuracy [73]. Afterwards, to address CNNs’ limited ability to model long-range dependencies, hybrid architectures that merge CNNs with RNN/LSTM layers were developed DeeperBind, DanQ, and iDeepS to predict DNA/RNA-binding protein interaction sites more accurately [63,65,74]. Although these CNN–RNN hybrids improved accuracy over CNNs alone, only modest gains were achieved, and the inability to learn very long sequences remained. To overcome these issues and to harness protein language models like ProTrans together with attention mechanisms, the transformer-based TransBind was proposed, achieving 97.68% accuracy and surpassing previous CNN and CNN–RNN models in binding-site prediction [75]. These advances in deep learning models can aid the discovery of previously unknown transcription factor binding regions, deepen understanding of gene-regulatory mechanisms, and help prioritize candidate regions efficiently during pre-experimental design.

3.2. Prediction of Expression Regulatory Regions and Estimation of Epigenome Characteristics

The importance of noncoding DNA regions that do not encode proteins has grown substantially in the regulation of gene expression. Although many analyses in genomics aim to elucidate the functions of noncoding DNA, predicting how noncoding regions

influence gene expression across diverse tissues and cell types remains challenging. Gene expression can be affected by the sequence features and epigenomic states of noncoding regions [76]. DeepChrome trains a CNN on labeled histone-modification marks within 10 kb around the TSS to predict gene expression strength. Compared with traditional ML approaches, DeepChrome achieves a higher AUC of 0.80, versus 0.66 for SVM and 0.59 for RF [77]. Subsequently, AttentiveChrome employed LSTMs to more precisely capture regulatory context [68]. In addition, because DNA methylation affects gene expression yet has often been assessed primarily by experimental means, CNN-based predictors such as DeepCpG and CpGenie were proposed to infer methylation status directly from sequence [78,79]. Building on these advances, Enformer has recently been developed as an integrated tool that simultaneously predicts gene expression and epigenomic features. Applying a transformer-based architecture, Enformer integrates distal regulatory information up to 100 kb away to model enhancer–gene links and predict expression [19]. Enformer outperforms prior approaches and, owing to its extensibility, has served as a foundation for newer genome-function prediction models such as Borzoi [80]. However, because Enformer is trained primarily on genome sequence–level labels, its sequence-function prediction can be somewhat limited when generalizing to novel cellular contexts. Addressing this, EpiGePT, which integrates transcription-factor RNA-seq and 3D chromatin-contact information, shows superior sequence-function prediction compared with Enformer [81]. Overall, the latest regulatory region prediction models tend to integrate large-scale context and multi-omics data to further enhance predictive power.

3.3. Splicing Prediction

Alternative splicing is one of the key determinants of complexity in eukaryotic transcriptomes. It is implicated in a wide range of biological processes, including species-specific cell differentiation, telomere length maintenance, and diseases such as cancer and autism spectrum disorder. Nevertheless, identifying splicing signals and predicting their activity remains challenging [82,83]. To detect splicing alterations, CNN-based models such as DeepSplice and SpliceRover take splice-junction sequences as input and identify splice variants. DeepSplice achieves 96.1% accuracy, and SpliceRover likewise attains 96% accuracy [84,85]. SpliceAI was a CNN model trained on GENCODE mRNA transcript data that predicted splice junctions, and it identified a significant increase in variants accompanied by splicing alterations in patients with intellectual disability [86]. By detecting splicing changes and enabling base-level tracking of variant-induced alterations, these models hold promise for clinical interpretation.

3.4. Pre-miRNA Prediction and miRNA Target Prediction

miRNAs are post-transcriptional regulators that predominantly bind to the 3'UTR of target mRNAs to repress translation. Although tools such as miRDeep2 exist to predict miRNAs from RNA-seq data at the genome scale, their accuracy has been limited, necessitating experimental validation [87]. To address this, CNN-based models emerged deepMir, which takes RNA sequences as input, and miRDNN, which ingests secondary-structure features and minimum free energy (MFE) values [88,89]. Subsequently, to facilitate the use of secondary-structure information and achieve higher accuracy than CNNs alone, the transformer-based miRe2e was proposed [90]. miRe2e accepts raw genomic sequences and processes them through three components (structure prediction, MFE estimation, and a pre-miRNA classifier), achieving higher accuracy than deepMir. Predicting miRNA–target gene interactions is also a major challenge. For miRNA–gene interaction prediction, earlier tools relied on seed complementarity and minimum-free-energy rules, but suffered from high false-positive rates. Although deep learning integration in the miRNA field

is comparatively sparse, RNN-based DeepTarget and LSTM-driven approaches such as SG-LSTM-FRAME have been proposed [91,92]. SG-LSTM-FRAME embeds gene–miRNA sequences or topological information and is trained with verified interaction labels from resources like miRTarBase, attaining an AUC of 0.93. This represents a substantial advance for noncoding RNA–gene interaction prediction without direct experimentation.

3.5. Prediction of Non-Canonical DNA Structure

DNA commonly has a right-handed (B-DNA) but, under tension stress, can adopt left-handed conformation (Z-DNA). Z-DNA has recently attracted attention as diverse biological functions have been elucidated [93]. Z-DNA was previously predicted using thermodynamic methods, but accuracy was limited, and given their flip-on feature, they are also difficult to verify experimentally [94]. To predict Z-DNA with deep learning, the DeepZ model, which combines RNN and CNN architectures to compute the probability of Z-DNA transitions, has been proposed [20,67,95]. DeepZ assessed the probability of Z-DNA formation and achieved higher accuracy than existing thermodynamic models and nearly identical accuracy to ChIP-seq results. The advent of a more accurate predictive model prior to experimentation can improve the accessibility of Z-DNA research.

3.6. Biological Function Prediction in Genomic Sequence

In the early days, genomic sequence analysis tools included methods such as DANN that used DNNs to detect variants associated with diseases or phenotypic changes [96]. Recent models enabled by architectures such as transformers that can handle complex multi-omics inputs are moving beyond single-function analyses to attempt genome-wide functional annotation via pretraining and fine-tuning. A representative example is DNABERT. Through unsupervised pretraining with k-mer tokenization, DNABERT learns upstream and downstream context to capture the global functions of DNA [54]. DNABERT has outperformed state-of-the-art predictors such as BPNet and Enformer in tasks including promoter detection, TF-binding site prediction, motif analysis, splice-site identification, and variant effect prediction. As an unsupervised model, DNABERT can pretrain on unlabeled sequences and then yield broadly useful representations via downstream fine-tuning, making data acquisition more tractable than for Enformer. However, DNABERT, which directly utilizes the BERT architecture, also had limitations. To address transformers' high data/computation demands and short input-length constraints, the Hyena-based HyenaDNA was proposed [97]. Whereas DNABERT typically accepts inputs of <512 bp, HyenaDNA maintains single-nucleotide resolution for sequences up to 32 kb. Subsequently, another transformer-based model for genome functional analysis, the Nucleotide Transformer, was proposed to improve the limitations of DNABERT. Using unsupervised learning, it can detect genes, introns, coding and noncoding regions, and variants [98]. After fine-tuning, Nucleotide Transformer achieved a higher Matthew's correlation coefficient than DNABERT, Enformer, and HyenaDNA. Taken together, deep learning models in genomics show a trend of improving accuracy in step with technological advances. However, improved long-range context modeling does not necessarily translate into stronger zero-shot performance [99]. Multi-omics integration and task-specific fine-tuning remain critical. The architectural progression from CNN, RNN, transformer/Hyena, combined with pretraining and multi-omics integration, continues to raise the resolution and accuracy of functional annotation, often down to single-base resolution. However, these gains typically come with increased demands for computational resources and larger training datasets (Table 3).

Table 3. Architectures and characteristics of major deep learning models proposed and applied to genomics research.

Function	Model	Architecture	Characteristics	Advantages	References
DNA/RNA binding protein binding sequence prediction	DeepBind	CNN	Protein–DNA/RNA binding prediction	- The first CNN-based protein binding predictor - simple to train	[53,64]
	iDeepS	CNN+LSTM	RNA–protein binding site prediction	- simple hybrid structure	[74]
	DanQ	CNN+LSTM	Regulatory feature prediction	- Strong performance on chromatin feature prediction	[65]
	TransBind	Transformer	Regulatory feature prediction	- Higher accuracy than CNN and RNN models	[75]
Epigenomic features	DeepChrome	CNN	Gene expression prediction from histone-modification profiles	- First CNN linking histone marks to gene activity - Simple and effective	[77]
	AttentiveChrome	LSTM+Attention	Gene expression prediction with attention over histone marks	- LSTM and attention improve interpretability - Identifies key histone marks driving expression	[68]
	CpGenie	CNN	DNA methylation prediction	- Strong performance on bulk methylation data	[79]
	DeepCpG	CNN	DNA methylation prediction	- Effective for single-cell methylation	[78]
	Enformer	Transformer	Gene expression, chromatin profile prediction	- Transformer captures long-range dependencies - Enables variant effect prediction	[19]
	EpiGePT	Transformer	Mapping gene expression levels from epigenetic marks	- Enables variant effect prediction - Outperforms CNN-based models	[81]
Splicing	DeepSplice	CNN	Alternative splicing detection	- Easy to train	[84]
	SpliceRover	CNN	Splice site detection	- High accuracy CNN for splice junctions	[85]
	SpliceAI	CNN	Quickly rank splicing effects by sequence	- Direct learning of long-distance context	[86]
miRNA	miRe2e	Transformer	Pre-miRNA prediction (Sequence, structure, MFE)	- Specialized for pre-miRNA prediction	[90]
	SG-LSTM-FRAME	LSTM	miRNA–mRNA interaction prediction	- High accuracy in miRNA target prediction	[92]
Non-canonical DNA structure	DeepZ	CNN+RNN	Z-DNA forming potential	- Specialized for left-handed Z-DNA detection	[67]
Genome function from sequence	DANN	DNN	Using a similar set of CADDs, but with weights automatically learned by deep learning.	- Wide application - Higher accuracy than existing machine learning	[96]
	BPNet	CNN	Base-resolution chromatin profiling	- Base-level motif discovery - Biologically validated	[73]
	DNABERT	Transformer	k-mer BERT pretraining and fine-tuning for tasks	- Flexible for multiple downstream tasks such as promoter, enhancer, and splicing site	[54]
	Borzoi	Transformer	Regulatory prediction, Using Enformer architecture	- Successor to Enformer with improved accuracy - Better representation of distal interactions	[80]
	HyenaDNA	Hyena	Long-range sequence prediction	- Efficiently models 10~35kb genomic contexts	[97]
	Nucleotide Transformer	Transformer	Masked sequence prediction	- Learns rich sequence representations without labels	[98]

4. Generative Algorithms Are Also Used in Genomics and Genetics to Advance Sequence Design and Data Augmentation

With the onset of the 2020s, AI moved beyond primarily classificational tasks to generative models that synthesize and reconstruct content from learned representations in response to user prompts. These generative algorithms construct images or sentences based on input prompts and conduct further learning based on the input data [100]. More recently, generative algorithms have begun to influence biology and genomics, spurring efforts to reconstruct genomic elements and to recapitulate characteristic genomic features.

Generative algorithms are increasingly aiding the conduct of genomics research. As a representative example, conversational AI like GPT has been used to analyze and categorize genomics research articles, enabling the inference of about 80% of organismal traits and the extraction of roughly 61% of marker–trait associations [101]. Similarly, large language models (LLMs) such as Gemini and Grok are also utilized in fields such as clinical reasoning [102]. Generative adversarial networks (GANs) learn from random noise via generator–discriminator competition to synthesize sequences and images, and they can model and generate characteristic patterns of genomic sequences [103,104]. Generative models such as GANs can synthesize additional training data and are used in genomics to augment datasets by generating synthetic genomes and transcriptomes [105]. There have been attempts to utilize GANs to generate specific DNA sequences and replicate genomic features [106]. As genomic datasets expand, an individual's gene and whole-genome sequences constitute sensitive personal information; together with increasingly stringent privacy protections, this hampers access to the DNA sequence data required for training. Consequently, the amount of data that can be leveraged for AI training is limited, which may degrade model performance [107]. To address these limitations, researchers have sought to recapitulate genomic features and generate artificial genomes. For example, Burak et al. trained GANs and Boltzmann machines on data from the data of 1000 Genomes Project and approximately 2000 Estonian genomes, producing high-quality artificial genomes that preserve key properties of real genomes. These artificial genome datasets can serve as valuable additional training data for downstream machine-learning studies [108]. Generative models such as GANs have captured genomic structure with considerable precision, opening new avenues for genomics research that directly generate genomes. Nonetheless, important limitations remain; for example, GANs still struggle to read and analyze longer sequences in a single pass.

The genome can be viewed as a collection of nucleotide sequences over a four-letter alphabet. This perspective has motivated efforts to directly apply LLM architectures to genomic data. LLM-based methods are now used across many areas of genomics to infer disease risk, predict the phenotypic effects of genetic variants, and inform personalized diagnostic medicine [109–111]. A representative architecture is the transformer, which underpins systems such as ChatGPT. An expanding body of genomic research now leverages transformer-based models. For example, DNABERT trains on large-scale sequence data with deep transformer stacks to predict regulatory elements in the genome, including promoters, splice sites, and transcription factor binding sites [54]. Beyond these examples, transformers have been applied across diverse models for genomic data analysis. In bacteria, transformer-based models identify transcription start sites, translation initiation sites, and DNA methylation loci to provide insight into transcriptional processes, and in human functional genomics, they are used to build models that infer phenotypes directly from DNA sequence [98,112–114]. However, in the field of genome generation, transformers require substantial computation and struggle with very long sequences [115]. This means that the basic unit for learning and constructing DNA sequences can be shortened, making it difficult to perfectly reflect the characteristics of the genome, which serves as a

catalyst for the need for next-generation architectures. Nguyen et al. present EVO, a truly genome-generating AI. The EVO genome-generating algorithm is based on StripedHyena, considered a next-generation LLM architecture, and addresses the problem of the transformer's difficulty in processing long information due to its token size [72,116]. EVO is designed to conduct large-scale token unit learning of 2K size, using a multimodal algorithm that predicts multiple factors rather than a single model, such as proteins, regulatory regions, DNA, and RNA, to reconstruct the features of the human genome [116,117]. These technological advancements mark the beginning of a path toward biological design. As technology advances, it becomes harder for humans to understand how AI works. In fact, there are even academic fields dedicated to figuring this out [118,119]. However, these advances offer many new perspectives on the current understanding of organisms in the field of genomics and may even lead to solutions to potential problems in genome design that even humans cannot identify.

Taken together, these perspectives and developments suggest that, in the near future, it may be possible to develop algorithms that generate genomic sequences from phenotypic information or, conversely, predict phenotypes from sequence data. Current genome design using generative AI is limited to how closely it infers existing sequences and how well it represents specific phenotypes. However, this demonstrates the potential for ushering in an era of biological design, where humans design the genomes and characteristics of living organisms.

5. Conclusions

Genomics has entered an era where AI models not only classify signals but also infer mechanisms and increasingly generate biological sequences. Classical ML provided the first scalable tools for genotype–phenotype mapping; deep learning then expanded the solution space, with CNNs uncovering local sequence rules, RNN/LSTMs capturing positional and temporal dependencies, and transformers learning distal regulatory grammar across hundreds of kilobases. Foundation models such as DNABERT, Enformer, and Nucleotide Transformer demonstrate that pretraining on unlabeled sequence augmented by multi-omics yields transferable representations for promoters, TF binding, splicing, methylation, and variant-effect prediction. These advances in AI tools have substantially propelled genomics, enabling the capture of features that were difficult to infer using traditional methods and tools. Moreover, the progress of generative models suggests that in the near future, we stand at the threshold of moving beyond genome analysis to genome design, enabling controllable sequence design under appropriate safeguards. At present, generative models such as EVO can replicate or design CRISPR guide sequences and functional gene sequences, and GANs are being used to prototype human artificial genomes. In the near future, algorithms may emerge that construct whole, functional genomes conditioned on phenotype or predict phenotypes directly from genotype.

Author Contributions: Conceptualization, D.H.L. and H.-S.K.; investigation, D.H.L., Y.J.L., H.-S.J., H.-Y.R., G.-r.J. and S.-W.K.; writing—original draft preparation, D.H.L. and E.G.P.; writing—review and editing, D.H.L., E.G.P., Y.J.L. and S.-W.K.; supervision, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This review received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Hill, J.E.; Harris, C.; Clegg, A. Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Res. Synth. Methods* **2024**, *15*, 347–353. [\[CrossRef\]](#)
- Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murgesan, K.; Mattei, N.; Rossi, F.; et al. Thinking fast and slow in AI. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 15042–15046. [\[CrossRef\]](#)
- Kwon, C. AI and the future of architecture: A Smart secretary, revolutionary tool, or a cause for concern? *Int. J. Sustain. Build. Technol. Urban Dev.* **2023**, *14*, 128–131. [\[CrossRef\]](#)
- Torrente, M.; Sousa, P.A.; Hernández, R.; Blanco, M.; Calvo, V.; Collazo, A.; Guerreiro, G.R.; Núñez, B.; Pimentao, J.; Sánchez, J.C.; et al. An artificial intelligence-based tool for data analysis and prognosis in cancer patients: Results from the clarify study. *Cancers* **2022**, *14*, 4041. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lepakshi, V.A. Machine learning and deep learning based AI tools for development of diagnostic tools. In *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection*; Academic Press: Cambridge, MA, USA, 2022; pp. 399–420. [\[CrossRef\]](#)
- Ghaffar Nia, N.; Kaplanoglu, E.; Nasab, A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discov. Artif. Intell.* **2023**, *3*, 5. [\[CrossRef\]](#)
- Boulesteix, A.L.; Wright, M. Artificial intelligence in genomics. *Hum. Genet.* **2022**, *141*, 1449–1450. [\[CrossRef\]](#)
- Wang, D.; Huang, G.B. Protein sequence classification using extreme learning machine. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; IEEE: New York, NY, USA, 2005; pp. 1406–1411. [\[CrossRef\]](#)
- Jeck, W.R.; Siebold, A.P.; Sharpless, N.E. Review: A meta-analysis of GWAS and age-associated diseases. *Aging Cell* **2012**, *11*, 727–731. [\[CrossRef\]](#)
- Franke, L.; Jansen, R.C. eQTL analysis in humans. *Cardiovasc. Genom. Methods Protoc.* **2009**, *573*, 311–328. [\[CrossRef\]](#)
- Dixit, P.; Prajapati, G.I. Machine learning in bioinformatics: A novel approach for DNA sequencing. In Proceedings of the 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, India, 21–22 February 2015; IEEE: New York, NY, USA, 2015; pp. 41–47. [\[CrossRef\]](#)
- Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. Available online: <https://www.nature.com/articles/s41586-021-03819-2> (accessed on 5 August 2025). [\[CrossRef\]](#) [\[PubMed\]](#)
- Caudai, C.; Calizia, A.; Geraci, F.; Le Pera, L.; Morea, V.; Salerno, E.; Via, A.; Colombo, T. AI applications in functional genomics. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 5762–5790. [\[CrossRef\]](#) [\[PubMed\]](#)
- Keshavarzi Arshadi, A.; Webb, J.; Salem, M.; Crus, E.; Calad-Thomson, S.; Ghadirian, N.; Collins, J.; Diez-Cecilia, E.; Kelly, B.; Goodarzi, H.; et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. *Front. Artif. Intell.* **2020**, *3*, 65. [\[CrossRef\]](#)
- Karmakar, T. Nobel Prize in Chemistry 2024: Computational Protein Design and Structure Prediction. *Resonance* **2025**, *30*, 649–662. [\[CrossRef\]](#)
- Nguyen, E.; Poli, M.; Durrant, M.G.; Kang, B.; Katrekar, D.; Li, D.B.; Bartie, L.J.; Thomas, R.W.; King, S.H.; Brixi, G.; et al. Sequence modeling and design from molecular to genome scale with Evo. *Science* **2024**, *386*, eado9336. [\[CrossRef\]](#)
- Logsdon, G.A.; Vollger, M.R.; Eichler, E.E. Long-Read Human Genome Sequencing and Its Applications. *Nat. Rev. Genet.* **2020**, *21*, 597–614. Available online: <https://www.nature.com/articles/s41576-020-0236-x> (accessed on 5 August 2025). [\[CrossRef\]](#) [\[PubMed\]](#)
- Jiang, S.; Mortazavi, A. Integrating ChIP-seq with other functional genomics data. *Brief. Funct. Genom.* **2018**, *17*, 104–115. [\[CrossRef\]](#) [\[PubMed\]](#)
- Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J.R.; Grabska-Barwinska, A.; Taylor, K.R.; Assael, Y.; Jumper, J.; Kohli, P.; Kelley, D.R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **2021**, *18*, 1196–1203. Available online: <https://www.nature.com/articles/s41592-021-01252-x> (accessed on 5 August 2025). [\[CrossRef\]](#)
- Beknazarov, N.; Poptsova, M. DeepZ: A deep learning approach for Z-DNA prediction. In *Z-DNA: Methods and Protocols*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 217–226. [\[CrossRef\]](#)
- Ambrish, G.; Ganesh, B.; Ganesh, A.; Srinivas, C.; Dhanraj; Mensinkal, K. Logistic regression technique for prediction of cardiovascular disease. *Glob. Transit. Proc.* **2022**, *3*, 127–130.
- Salman, H.A.; Kalakech, A.; Steiti, A. Random forest algorithm overview. *Babylon. J. Mach. Learn.* **2024**, *2024*, 69–79. [\[CrossRef\]](#)
- Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurobot.* **2013**, *7*, 21. [\[CrossRef\]](#)
- Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 218. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yang, F.J. An implementation of naive bayes classifier. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 12–14 December 2018; IEEE: New York, NY, USA, 2018; pp. 301–306. [\[CrossRef\]](#)

26. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 101–121. [\[CrossRef\]](#)
27. Sambo, F.; Trifoglio, E.; Di Camillo, B.; Toffolo, G.M.; Cobelli, C. Bag of Naïve Bayes: Biomarker selection and classification from genome-wide SNP data. *BMC Bioinform.* **2012**, *13* (Suppl. S14), S2. [\[CrossRef\]](#)
28. Ban, H.J.; Heo, J.Y.; Oh, K.S.; Park, K.J. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *BMC Genet.* **2010**, *11*, 26. [\[CrossRef\]](#)
29. Bagheri, M.; Miraie-Ashtiani, R.; Moradi-Shahrbabak, M.; Nejati-Javaremi, A.; Pakdel, A.; von Borstel, U.U.; Pimentel, E.C.G.; König, S. Selective genotyping and logistic regression analyses to identify favorable SNP-genotypes for clinical mastitis and production traits in Holstein dairy cattle. *Livest. Sci.* **2013**, *151*, 140–151. [\[CrossRef\]](#)
30. Briggs, F.; Ramsay, P.P.; Madden, E.; Norris, J.M.; Holers, V.M.; Mikuls, T.R.; Sokka, T.; Seldin, M.F.; Gregersen, P.K.; Criswell, L.A.; et al. Supervised Machine Learning and Logistic Regression Identifies Novel Epistatic Risk Factors with PTPN22 for Rheumatoid Arthritis. *Genes Immun.* **2010**, *11*, 199–208. Available online: <https://www.nature.com/articles/gene2009110> (accessed on 12 August 2025). [\[CrossRef\]](#)
31. Araújo, G.S.; Souza, M.R.B.; Oliveira, J.R.M.; Costa, I.G. Random forest and gene networks for association of SNPs to Alzheimer's disease. In *Brazilian Symposium on Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 104–115. [\[CrossRef\]](#)
32. Sun, Y.V. Multigenic modeling of complex disease by random forests. *Adv. Genet.* **2010**, *72*, 73–99. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Schwender, H. Imputing missing genotypes with weighted k nearest neighbors. *J. Toxicol. Environ. Health A* **2012**, *75*, 438–446. [\[CrossRef\]](#)
34. Yang, C.H.; Weng, Z.J.; Chuang, L.Y.; Yang, C.S. Identification of SNP-SNP interaction for chronic dialysis patients. *Comput. Biol. Med.* **2017**, *83*, 94–101. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Parry, R.; Jones, W.; Stokes, T.H.; Phan, J.H.; Moffitt, R.A.; Fang, H.; Shi, L.; Oberthuer, A.; Fischer, M.; Tong, W.; et al. k-Nearest Neighbor Models for Microarray Gene Expression Analysis and Clinical Outcome Prediction. *Pharmacogenomics J.* **2010**, *10*, 292–309. Available online: <https://www.nature.com/articles/tpj201056> (accessed on 12 August 2025). [\[CrossRef\]](#)
36. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [\[CrossRef\]](#)
37. Li, Y.; Zou, Z.; Gao, Z.; Wang, Y.; Xiao, M.; Xu, C.; Jiang, G.; Wang, H.; Jin, L.; Wang, J.; et al. Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. *Cancer Med.* **2022**, *11*, 4469–4478. [\[CrossRef\]](#)
38. Enoma, D.; Osamor, V.C.; Olubanke, O. Extreme gradient boosting machine learning algorithm identifies genome-wide genetic variants in prostate cancer risk prediction. *BioRxiv* **2023**. [\[CrossRef\]](#)
39. Ahmed, H.; Soliman, H.; Elmogy, M. Early detection of Alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree. *Comput. Biol. Med.* **2022**, *146*, 105622. [\[CrossRef\]](#)
40. Wei, W.; Visweswaran, S.; Cooper, G.F. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 370–375. [\[CrossRef\]](#)
41. Chuang, L.Y.; Wu, K.C.; Chang, H.W.; Yang, C.H. Support vector machine-based prediction for oral cancer using four snps in DNA repair genes. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, China, 16–18 March 2011.
42. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51.
43. Liu, J.; Li, J.; Wang, H.; Yan, J. Application of deep learning in genomics. *Sci. China Life Sci.* **2020**, *63*, 1860–1878. [\[CrossRef\]](#)
44. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [\[CrossRef\]](#)
45. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural. Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
47. Yang, Z.; Zeng, X.; Zhao, Y.; Chen, R. AlphaFold2 and Its Applications in the Fields of Biology and Medicine. *Signal Transduct. Targeted Ther.* **2023**, *8*, 115. Available online: <https://www.nature.com/articles/s41392-023-01381-z> (accessed on 15 August 2025). [\[CrossRef\]](#)
48. Rajoub, B. Supervised and unsupervised learning. In *Biomedical Signal Processing and Artificial Intelligence in Healthcare*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 51–89. [\[CrossRef\]](#)
49. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. Available online: <https://www.nature.com/articles/nrg3920> (accessed on 20 August 2025). [\[CrossRef\]](#)
50. Saravanan, R.; Sujatha, P.P. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 14–15 June 2018; IEEE: New York, NY, USA, 2018; pp. 945–949. [\[CrossRef\]](#)

51. Omta, W.A.; von Heesbeen, R.G.; Egan, D.A. Combining supervised and unsupervised machine learning methods for phenotypic functional genomics screening. *SLAS Discov.* **2020**, *25*, 655–664. [CrossRef]
52. Ang, J.C.; Mirzal, A.; Haron, H.; Hamed, H.N.A. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *TCBB* **2015**, *13*, 971–989. [CrossRef] [PubMed]
53. Alipanahi, B.; Delong, A.; Weirauch, M.T.; Frey, B.J. Predicting the Sequence Specificities of DNA-and RNA-Binding Proteins by Deep Learning. *Nat. Biotechnol.* **2015**, *33*, 831–838. Available online: <https://www.nature.com/articles/nbt.3300> (accessed on 15 August 2025). [CrossRef]
54. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [CrossRef]
55. Date, Y.; Kikuchi, J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. *Anal. Chem.* **2018**, *90*, 1805–1810. [CrossRef] [PubMed]
56. Koutsoukas, A.; Monaghan, K.J.; Li, X.; Huan, J. Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **2017**, *9*, 42. [CrossRef]
57. Cheng, Y.; Wang, D.; Zhou, P.; Zhang, T. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Process. Mag.* **2018**, *35*, 126–136. [CrossRef]
58. Ye, J.; Wang, S.; Yang, X.; Tang, X. Gene prediction of aging-related diseases based on DNN and Mashupp. *BMC Bioinform.* **2021**, *22*, 597. [CrossRef] [PubMed]
59. Kim, S.G.; Harwani, M.; Grama, A.; Chaterji, S. EP-DNN: A Deep Neural Network-Based Global Enhancer Prediction Algorithm. *Sci. Rep.* **2016**, *6*, 38433. Available online: <https://www.nature.com/articles/srep38433> (accessed on 20 August 2025). [CrossRef]
60. Ketkar, N.; Moolayil, J. Convolutional neural networks. In *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 197–242. [CrossRef]
61. Yang, J.; Li, J. Application of deep convolution neural network. In Proceedings of the 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 15–17 December 2017; IEEE: New York, NY, USA, 2017; pp. 229–232. [CrossRef]
62. Arkin, E.; Yadikar, N.; Xu, X.; Aysa, A.; Ubul, K. A survey: Object detection methods from CNN to transformer. *Multim. Tools Appl.* **2023**, *82*, 21353–21383. [CrossRef]
63. Hassanzadeh, H.R.; Wang, M.D. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; IEEE: New York, NY, USA, 2016; pp. 178–183. [CrossRef]
64. Kelley, D.R.; Snoek, J.; Rinn, J.L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **2016**, *26*, 990–999. [CrossRef]
65. Quang, D.; Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **2016**, *44*, e107. [CrossRef]
66. Tavakoli, N. Modeling genome data using bidirectional LSTM. In Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, 15–19 July 2019; IEEE: New York, NY, USA, 2019; pp. 183–188. [CrossRef]
67. Beknazarov, N.; Jin, S.; Poptsova, M. Deep learning approach for predicting functional Z-DNA regions using omics data. *Sci. Rep.* **2020**, *10*, 19134. Available online: <https://www.nature.com/articles/s41598-020-76203-1> (accessed on 20 August 2025). [CrossRef] [PubMed]
68. Singh, R.; Lanchantin, J.; Sekhon, A.; Qi, Y. Attend and predict: Understanding gene regulation by selective attention on chromatin. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6785–6795.
69. Jawahar, G.; Sagot, B.; Seddah, D. What does BERT learn about the structure of language? In Proceedings of the ACL 2019–57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 29–31 July 2019.
70. Consens, M.E.; Dufault, C.; Wainberg, M.; Forster, D.; Karimzadeh, M.; Goodarzi, H.; Theis, F.J.; Moses, A.; Wang, B. Transformers and Genome Language Models. *Nat. Mach. Intell.* **2025**, *7*, 346–362. Available online: <https://www.nature.com/articles/s42256-025-01007-9> (accessed on 20 August 2025). [CrossRef]
71. Della Libera, L.; Subakan, C.; Ravanelli, M.; Cornell, S.; Lepoutre, F.; Grondin, F. Resource-efficient separation transformer. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 761–765. [CrossRef]
72. Poli, M.; Massaroli, S.; Nguyen, E.; Fu, D.Y.; Dao, T.; Baccus, S.; Bengio, Y.; Ermon, S.; Re, C. Hyena hierarchy: Towards larger convolutional language models. *PMLR* **2023**, *202*, 28043–28078.
73. Avsec, Ž.; Weilert, M.; Shrikumar, A.; Krueger, S.; Alexandari, A.; Dalal, K.; Fropf, R.; McAnany, C.; Gagneur, J.; Kundaje, A. Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax. *Nat. Genet.* **2021**, *53*, 354–366. Available online: <https://www.nature.com/articles/s41588-021-00782-6> (accessed on 20 August 2025). [CrossRef] [PubMed]

74. Pan, X.; Rijnbeek, P.; Yan, J.; Shen, H.B. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genom.* **2018**, *19*, 511. [CrossRef] [PubMed]
75. Tahmid, M.T.; Hasan, A.M.; Bayzid, M.S. TransBind Allows Precise Detection of DNA-Binding Proteins and Residues Using Language Models and Deep Learning. *Commun. Biol.* **2025**, *8*, 568. Available online: <https://www.nature.com/articles/s42003-025-07534-w> (accessed on 20 August 2025). [CrossRef] [PubMed]
76. Barrett, L.W.; Fletcher, S.; Wilton, S.D. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.* **2012**, *69*, 3613–3634. [CrossRef]
77. Singh, R.; Lanchantin, J.; Robins, G.; Qi, Y. DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **2016**, *32*, i639–i648. [CrossRef]
78. Angermueller, C.; Lee, H.J.; Reik, W.; Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* **2017**, *18*, 67. [CrossRef]
79. Zeng, H.; Gifford, D.K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res.* **2017**, *45*, e99. [CrossRef]
80. Linder, J.; Srivastava, D.; Yuan, H.; Agarwal, V.; Kelley, D.R. Predicting RNA-Seq Coverage from DNA Sequence as a Unifying Model of Gene Regulation. *Nat. Genet.* **2025**, *57*, 949–961. Available online: <https://www.nature.com/articles/s41588-024-02053-6> (accessed on 20 August 2025). [CrossRef] [PubMed]
81. Gao, Z.; Zeng, W.; Jiang, R.; Wong, W.H. EpiGePT: A pretrained transformer-based language model for context-specific human epigenomics. *Genome Biol.* **2024**, *25*, 310. [CrossRef]
82. Marasco, L.E.; Kornblihtt, A.R. The Physiology of Alternative Splicing. *Nat. Rev. Mol. Cell Biol.* **2023**, *24*, 242–254. Available online: <https://www.nature.com/articles/s41580-022-00545-z> (accessed on 20 August 2025). [CrossRef] [PubMed]
83. Xu, C.; Kornblihtt, A.R. Reference-informed prediction of alternative splicing and splicing-altering mutations from sequences. *Genome Res.* **2024**, *34*, 1052–1065. [CrossRef] [PubMed]
84. Zhang, Y.; Liu, X.; MacLeod, J.N.; Liu, J. DeepSplice: Deep classification of novel splice junctions revealed by RNA-seq. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; IEEE: New York, NY, USA, 2016; pp. 330–333. [CrossRef]
85. Zuallaert, J.; Godin, F.; Kim, M.; Soete, A.; Saeys, Y.; De Neve, W. SpliceRover: Interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **2018**, *34*, 4180–4188. [CrossRef]
86. Jaganathan, K.; Panagiotopoulou, S.K.; McRae, J.F.; Darbandi, S.F.; Knowles, D.; Li, Y.I.; Kosmicki, J.A.; Arbelaez, J.; Cui, W.; Schwartz, G.B.; et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **2019**, *176*, 535–548. [CrossRef]
87. Friedländer, M.R.; Mackowiak, S.D.; Li, N.; Chen, W.; Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **2012**, *40*, 37–52. [CrossRef]
88. Tang, X.; Sun, Y. Fast and accurate microRNA search using CNN. *BMC Bioinform.* **2019**, *20* (Suppl. S23), 646. [CrossRef]
89. Yones, C.; Raad, J.; Bugnon, L.A.; Milone, D.H.; Stegmayer, G. High precision in microRNA prediction: A novel genome-wide approach with convolutional deep residual networks. *Comput. Biol. Med.* **2021**, *134*, 104448. [CrossRef]
90. Raad, J.; Bugnon, L.A.; Milone, D.H.; Stegmayer, G. miRe2e: A full end-to-end deep model based on transformers for prediction of pre-miRNAs. *Bioinformatics* **2022**, *38*, 1191–1197. [CrossRef] [PubMed]
91. Lee, B.; Baek, J.; Park, S.; Yoon, S. deepTarget: End-to-end learning framework for microRNA target prediction using deep recurrent neural networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Seattle, WA, USA, 2–5 October 2016. [CrossRef]
92. Xie, W.; Luo, J.; Pan, C.; Liu, Y. SG-LSTM-FRAME: A computational frame using sequence and geometrical information via LSTM to predict miRNA–gene associations. *Brief. Bioinform.* **2021**, *22*, 2032–2042. [CrossRef]
93. Herbert, A. Z-DNA and Z-RNA in Human Disease. *Commun. Biol.* **2019**, *2*, 7. Available online: <https://www.nature.com/articles/s42003-018-0237-x> (accessed on 22 August 2025). [CrossRef]
94. Czarny, R.S.; Ho, P.S. Thermogenomic analysis of left-handed Z-DNA propensities in genomes. In *Z-DNA: Methods and Protocols*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 195–215. [CrossRef]
95. Umerenkov, D.; Kokh, V.; Herbert, A.; Poptsova, M. Generating Genomic Maps of Z-DNA with the Transformer Algorithm. In *Data Analysis and Optimization: In Honor of Boris Mirkin's 80th Birthday*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 363–376. [CrossRef]
96. Quang, D.; Chen, Y.; Xie, X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **2014**, *31*, 761–763. [CrossRef]
97. Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Wornow, M.; Birch-Sykes, C.; Massaroli, S.; Patel, A.; Rabideau, C.; Bengio, Y.; et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 43177–43201.

98. Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N.L.; Grzywaczewski, A.H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B.P.; Sirelkhatim, H.; et al. Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *Nat. Methods* **2025**, *22*, 287–297. Available online: <https://www.nature.com/articles/s41592-024-02523-z> (accessed on 22 August 2025). [[CrossRef](#)]
99. Benegas, G.; Eraslan, G.; Song, Y.S. Benchmarking DNA sequence models for causal regulatory variant prediction in human genetics. *bioRxiv* **2025**, preprint. [[CrossRef](#)] [[PubMed](#)]
100. Sengar, S.S.; Hasan, A.B.; Kumar, S.; Carroll, F. Generative artificial intelligence: A systematic review and applications. *Multimed. Tools Appl.* **2025**, *84*, 23661–23700. [[CrossRef](#)]
101. Poretsky, E.; Blake, V.C.; Andorf, C.M.; Sen, T.Z. Assessing the performance of generative artificial intelligence in retrieving information against manually curated genetic and genomic data. *Database* **2025**, *2025*, baaf011. [[CrossRef](#)]
102. Reyes-Rivera, J.; Molina, A.C.; Romero-Lorenzo, M.; Ali, S.; Gibson, C.; Saucedo, J.; Calandrelli, M.; Cruz, E.G.; Bahit, C.; Chi, G.; et al. Evaluating the Clinical Reasoning of GPT-4, Grok, and Gemini in Different Fields of Cardiology. *Circulation* **2024**, *150* (Suppl. S1), A4147550. [[CrossRef](#)]
103. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *CACM* **2020**, *63*, 139–144. [[CrossRef](#)]
104. Das, S.; Shi, X. Offspring GAN augments biased human genomic data. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Northbrook, IL, USA, 7–10 August 2022. [[CrossRef](#)]
105. Lacan, A.; Sebag, M.; Hanczar, B. GAN-based data augmentation for transcriptomics: Survey and comparative assessment. *Bioinformatics* **2023**, *39* (Suppl. S1), i111–i120. [[CrossRef](#)]
106. Killoran, N.; Lee, L.J.; Delong, A.; Duvenaud, D.; Frey, B.J. Generating and designing DNA with deep generative models. *arXiv* **2017**, arXiv:1712.06148. [[CrossRef](#)]
107. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.W. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **2023**, *123*, 8736–8780. [[CrossRef](#)] [[PubMed](#)]
108. Yelmen, B.; Decelle, A.; Ongaro, L.; Marnetto, D.; Tallec, C.; Montinaro, F.; Furtlehner, C.; Pagani, L.; Jay, F. Creating artificial human genomes using generative neural networks. *PLoS Genet.* **2021**, *17*, e1009303. [[CrossRef](#)] [[PubMed](#)]
109. Fallahpour, A.; Magnuson, A.; Gupta, P.; Ma, S.; Naimier, J.; Shah, A.; Duan, H.; Omar, I.; Goodarzi, H.; Maddison, C.J.; et al. BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model. *arXiv* **2025**, arXiv:2505.23579. [[CrossRef](#)]
110. Arulmurugan, A.; Eswa Sudhan, M.; Akshay Shravan, V.; Mazik, R.I.M.E. Enhanced Genetic Disorder Detection using LLM. In Proceedings of the 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), Erode, India, 11–13 June 2025; IEEE: New York, NY, USA, 2025; pp. 1360–1365. [[CrossRef](#)]
111. Liu, H.; Chen, S.; Wang, H. GenoAgent: A Baseline Method for LLM-Based Exploration of Gene Expression Data in Alignment with Bioinformaticians. Available online: <https://openreview.net/pdf?id=v7aeTmfGOu> (accessed on 4 August 2025).
112. Clauwaert, J.; Waegeman, W. Novel transformer networks for improved sequence labeling in genomics. *TCCB* **2020**, *19*, 97–106. [[CrossRef](#)]
113. Clauwaert, J.; Menschaert, G.; Waegeman, W. Explainability in transformer models for functional genomics. *Brief. Bioinform.* **2021**, *22*, bbab060. [[CrossRef](#)]
114. He, S.; Gao, B.; Sabnis, R.; Sun, Q. Nucleic transformer: Classifying dna sequences with self-attention and convolutions. *ACS Synth. Biol.* **2023**, *12*, 3205–3214. [[CrossRef](#)] [[PubMed](#)]
115. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [[CrossRef](#)] [[PubMed](#)]
116. Ralambomihanta, T.R.; Mohammadzadeh, S.; Islam, M.S.N.; Jabbour, W.; Liang, L. Scavenging hyena: Distilling transformers into long convolution models. *arXiv* **2024**, arXiv:2401.17574. [[CrossRef](#)]
117. Brixi, G.; Durrant, M.G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G.A.; King, S.H.; Li, D.B.; Merchant, A.T. Genome modeling and design across all domains of life with Evo 2. *BioRxiv* **2025**. [[CrossRef](#)]
118. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18. [[CrossRef](#)]
119. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.