Advancements in Gene Structure Prediction: Innovation and Prospects of Deep Learning Models Apply in Multi-species

Running title: Deep Learning in Gene Prediction Across Species

Tong Wang[1#*], Jing-Min Yang[2#*], Ting Xu[4†], Yuanyin Teng[5†], Yuqing Miao[1†] and Ming Wu[3*]

[1] Department of Biology, Duke University, Durham, NC 27708, the United States

[2] Hubei Bioinformatics & Molecular Imaging Key Laboratory, College of Life Science and Technology Huazhong University of Science and Technology, Wuhan 430074, China

[3] Gongli Hospital affiliated to Second Military Medical University, Shanghai 200135, China

[4] University of Massachusetts, Boston, MA 02125, the United States

[5] Institute of Hematology, Zhejiang University, Hangzhou 310003, China

#These authors contributed equally as the first authors: Tong Wang and Jing-Min Yang

†These authors contributed equally as the second authors: Ting Xu, Yuanyin Teng, and Yuqing Miao

*Correspondence: wangtongnelly@gmail.com, tong.wang@duke.edu (Tong Wang),
yangjingmin2021@163.com (Jing-Min Yang), mingwu819@163.com (Ming Wu)

ABSTRACT

In recent years, advancements in gene structure prediction have been significantly driven by the integration of deep learning technologies into bioinformatics. Transitioning from traditional thermodynamics and comparative genomics methods to modern deep learning-based models, such as CDSBERT, DNABERT, RNA-FM, and PlantRNA-FM prediction accuracy and generalization have seen remarkable improvements. These models, leveraging genome sequence data along with secondary and tertiary structure information, have facilitated diverse applications in studying gene functions across animals, plants, and humans. They also hold substantial potential for multi-application in early disease diagnosis, personalized treatment, and genomic evolution research. This review combines traditional gene structure prediction methods with advancements in deep learning, showcasing applications in functional region annotation, protein-RNA interactions, and cross-species genome analysis. It highlights their contributions to animal, plant, and human disease research while exploring future opportunities in cancer mutation prediction, RNA vaccine design, and CRISPR gene editing optimization. The review also emphasizes future directions, such as model refinement, multimodal integration, and global collaboration. By offering a concise overview and forward-looking insights, this article aims to serve as a fundamental resource and guide for advancing nucleic acid structure prediction research.

Keywords:

Gene Structure Prediction, Deep Learning, RNA language, multi-application, disease research

HIGHLIGHTS

- **Revolutionizing gene structure prediction with deep learning:**

  Advanced models like DNABERT, CDSBERT, and RNA-FM significantly enhance accuracy in functional region annotation, regulatory element identification, and cross-species genomic analysis.

- **Integrating multimodal data for comprehensive genomic insights:**

  Combining sequence, structure, and functional data through AI-driven models facilitates more accurate gene function predictions and enhances evolutionary and comparative genomic studies.

- **Transforming disease research and personalized medicine:**

  Deep learning-based gene structure prediction enables breakthroughs in early disease detection, RNA vaccine design, CRISPR optimization, and precision medicine.

INTRODUCTION

Before, the Human Genome Project and various organismal sequencing initiatives have produced an unprecedented abundance of biological data. The growing demand for data analysis and interpretation is being addressed by the evolving field of bioinformatics, applies computational methods to process, analyze, and interpret biological data [1]. The field of gene structure prediction has become a key research area in genetics, evolutionary biology, and disease research. Accurately predicting gene structure is essential for unraveling the complexity of gene function and regulation, with far-reaching implications across multiple fields. It plays a crucial role in advancing targeted medical therapies, deepening our understanding of evolutionary processes, and identifying genetic variants linked to diseases, ultimately driving innovations in healthcare and genomics research discovery. Within the field of gene structure prediction, despite the encouraging progress made in this field, there are still some challenges and bottlenecks. A key challenge lies in the intrinsic complexity of genome sequences, exhibiting significant variability and nonlinear patterns. This complexity poses a challenge for traditional machine learning algorithms, making it difficult to accurately discern underlying biological patterns.

Recent progress in data-driven approaches, particularly with the rise of machine learning and deep learning models, which have transformed gene structure prediction by enabling the analysis of large genomic datasets with unparalleled accuracy and efficiency [2]. With advancements in science and technology, the integration of machine learning and deep learning into genomic research has introduced new avenues for exploring the intricate relationship between gene sequences and their functional outcomes, thereby deepening our understanding of biological systems [3]. Moreover, deep learning-based models like Fusion AI can adapt to diverse genomic data types, enabling researchers to more accurately identify fusion genes and gain deeper insights into genomic breakage, making them versatile tools for applications across multiple kinds of species [4]. As these models continue to evolve, they will have the potential to deepen our understanding of gene regulation and expression, and paving the way for advancements in disease prediction, personalized medicine, genome sequencing, gene expression analysis, protein structure modeling, drug discovery, and disease diagnostics [5].

Addressing these challenges requires ongoing research to improve data quality, develop powerful algorithms, and enhance the interpretability of predictive models. To date, deep learning approaches have demonstrated significant advantages in integrating AI-driven techniques across various aspects of biological research. The utilized tools include DeepBind, DeepCpG, DeepGene, DeepFam, DeepLoc, DeepPath, ScanNet, and DeepVariant [6]. In addition, since the quality and completeness of genomic datasets can significantly affect the performance of prediction models, incomplete annotations and differences in gene structure between different databases are likely to lead to inaccurate predictions [7]. Therefore, deep learning model training necessitates a substantial volume of labeled data, that also poses challenges, especially in understudied species where genomic data may be scarce. For example, DRANetSplicer is designed to predict splice sites across different species by leveraging cross-organism validation [8]. To overcome the challenges, recent breakthroughs in deep learning have transformed gene structure prediction by harnessing large genomic datasets with advanced neural network architectures. By integrating primary sequence data with secondary and tertiary structural information, these models have significantly improved functional region annotation, regulatory element identification, and cross-species genomic comparisons. By overcoming some of the shortcomings and obstacles, the field of gene structure prediction can continue to advance, ultimately deepening the understanding of its impact on human health and disease treatment.

RESULTS

Predicting the structure of DNA and RNA is fundamental to understanding gene regulation, transcription, translation, and molecular interactions. DNA structure prediction focuses on identifying sequence properties such as helical stability, conformational dynamics, binding affinities, and methylation sites, which are crucial for locating regulatory elements like promoters and enhancers. RNA structure prediction, on the other hand, determines secondary and tertiary folding patterns that influence RNA stability, function, and interactions with proteins. Precise structural predictions offer valuable insights into gene function, RNA-protein interactions, and various applications, including RNA-based therapeutics such as mRNA vaccines. Over time, the DNA and RNA structures prediction methods have advanced from thermodynamic models based on free energy minimization to comparative sequence alignment techniques. While these traditional approaches provided foundational insights, they often struggled with computational limitations and species-specific variations. The emergence of machine learning and deep learning has revolutionized discipline, enabling the extraction of complex patterns from large genomic datasets with high accuracy. We collected Time-Lapse with selected representative software in gene structure prediction, which provides a systematic overview of the evolution of methodologies for DNA (Figure 1A) and RNA (Figure 1B) structure prediction, illustrating key advancements across different computational approaches. These methodologies are categorized into thermodynamic methods, comparative alignment methods, and deep learning methods, demonstrating their historical progression and impact on nucleic acid structural analysis. These tools and resources are available for the analysis of DNA/RNA structures. In a certain era, they have representative significance. The timeline presented (Figure 1) shows the progression from traditional alignment and thermodynamic-based methods to modern machine learning and deep learning techniques, underscoring their increasing sophistication and predictive power. These computational tools serve as critical resources for genomic research, RNA structural modeling, functional annotation, and biomedical applications, facilitating advancements in cancer research, gene regulation studies, and molecular diagnostics. Understanding this progression highlights how technological advancements have transformed our ability to decode nucleic acid structures, paving the way for more precise applications in functional genomics, disease research, and biotechnological innovations.

Traditional Methods for Gene Sequence and Structural Analysis

Early Methods

Early methods for gene sequence and structure analysis revolved around basic sequence alignment techniques and manual curation of genomic data and thermodynamic models. Although very fundamental, these methods are fundamentally constrained by their dependence on direct sequence comparisons and the manual curation of genetic features. The introduction of sequence alignment algorithms marked a major advance in the field. However, these methods are computationally intensive and often require a lot of time and expertise to accurately interpret the results. The reliance on manual curation means that many genomic sequences are still poorly annotated, resulting in gaps in deciphering gene function and architecture. Moreover, early methods lacked the ability to process and manage the vast volumes of data produced by modern sequencing technologies, necessitating the advancement of more sophisticated computational tools and models capable of efficiently processing and analyzing genomic data [9,10].

Thermodynamic Models

(e.g., Gene Structure and RNA Secondary Structure Prediction)

Thermodynamic Models were used for both DNA and RNA structure analysis. Thermodynamic models, particularly those used for predicting RNA secondary structures, represent a significant advancement in the analysis of gene sequences. These models apply thermodynamic principles to determine the most stable RNA configurations based on their nucleotide sequences. The free energy minimization approach allows researchers to infer potential secondary structures by calculating the stability of various folding patterns. Thermodynamic models for DNA structure analysis are methods used to predict and evaluate DNA structures based on thermodynamic principles, such as free energy minimization, base-pair stability, and stacking interactions. These models rely on experimentally derived parameters to understand DNA folding, hybridization, and secondary structure formation

Tools such as OligoAnalyzer (Predicts DNA duplex stability and melting temperature), Mfold (Predicts DNA secondary structures based on free energy minimization), UNAFold (Integrates various thermodynamic models for DNA and RNA analysis), ViennaRNA (RNA secondary structure prediction and comparison) and RNAfold (Processes individual RNA sequences and calculates their minimum free energy structures) have become standard in the field, enabling researchers to visualize and predict RNA structures efficiently. However, these models are not without limitations; they often rely on idealized conditions that may not accurately reflect the complexities of biological environments. Furthermore, the prediction accuracy can be influenced by the presence of non-canonical base pairs and the dynamic nature of RNA folding, which can complicate the interpretation of results [9].

Comparative Genomics and Sequence Alignment Methods

Comparative genomics has become a valuable approach for exploring evolutionary relationships and functional genomics [11]. By analyzing and comparing genomes across species, researchers can detect conserved sequences and deduce gene functions through homology-based methods such as BLAST [12]. Other sequence alignment methods, including multiple sequence alignment (MSA) algorithms, have been developed to facilitate these comparisons.

The first truly practical approach to MSA was developed by 1987 by Needleman–Wunsch alignment [13]. Widely used global sequence alignment algorithms include Optimal and Heuristic-based methods such as AlignMe, Needleman-Wunsch, GLASS, WABA, AVID, and CHAOS. In contrast, the most utilized local sequence alignment algorithms include Smith-Waterman, FASTA, BLAST, BLASTZ, PatternHunter, YASS, LAMBDA, USearch, LAST, and ALLAlign [14]. Local sequence alignments are specifically designed to identify matching subregions within two sequences, typically requiring less computational time compared to global alignment algorithms [15]. The Smith-Waterman algorithm is built on a technique known as Dynamic Programming. To address the limitations of optimal algorithms, heuristic algorithms were later developed as one of the more efficient alternatives. The earliest heuristic algorithm, FASTA (Fast-All), was developed by Lipman and Pearson as an efficient approach to sequence alignment [15]. BLAST utilizes a look-up table to detect seed matches, making it faster than FASTA [16].

TABLE 1. (Modified from [17]) provides an overview of multiple sequence alignment (MSA) tools, comparing their input formats, output formats, sequence types, methods, and server availability. It includes widely used tools like CLUSTAL OMEGA, MAFFT, MUSCLE, KALIGN, RETALIGN, and PROBCONS, each employing different alignment strategies such as progressive, iterative, and probabilistic consistency-based methods. These tools support protein, DNA, and RNA sequence alignment, facilitating accurate evolutionary and functional genomics studies. This table also includes links to their respective online servers for easy access.

This phylogenetic analysis can take weeks to complete, even when performed on High-Performance Computing (HPC) systems [18]. Multiple Sequence Alignment (MSA) is a fundamental technique in comparative genomics and sequence alignment. It is widely utilized in phylogenetic analysis to construct evolutionary trees, allowing researchers to determine evolutionary connections among homologous genes. By aligning three or more biological sequences. Tools like Clustal Omega, MAFFT, MUSCLE, and MEGA enable simultaneous alignment of multiple sequences, offering insights into evolutionary conservation and divergence. However, the accuracy of MSA relies heavily on the quality of input data and the availability of computational resources. In addition, the presence of highly divergent sequences complicates alignment accuracy, leading to potential misinterpretations of evolutionary relationships. Therefore, improving algorithms to handle large data sets and accounting for the complexity of genomic evolution has become a new challenge [17].

Gene structure prediction can be broadly categorized into RNA-Seq-based methods, homology-based techniques, and either individual prediction models or integrative methods that combine multiple approaches for enhanced accuracy. Integrative methods enhance prediction accuracy by merging multiple approaches or selecting the most reliable individual prediction outcomes; GINGER uses Next flow to predict gene structure, which enhances prediction accuracy at the exon level while optimizing computational resources for greater efficiency and effectiveness. It is particularly well-suited for species with highly complex gene structures [19].

Limitations of Traditional Methods

Although great progress has been made in gene sequence and structure analysis methods in methodological research in recent years, there are still many limitations. A major challenge is the strong reliance on annotation data. The reliance on annotation data for analysis may lead to analytical bias, which will limitation restricts the discovery of novel genes and their respective functions. Furthermore, the expandability of numerous conventional approaches is frequently constrained by the intrinsic complexity of their algorithms, making it difficult to extend the analysis to multiple species or large populations, and there are difficulties in computational efficiency, especially when applied to large-scale genomic data sets. As genomic data and protein structure analysis continues to grow exponentially, the need for more efficient, scalable, and automated methods becomes increasingly important, and with the improvement of old analysis methods and the emergence of new methods, people's understanding of genomics and its applications in medicine, protein structure, disease prediction and biotechnology will surely be further improved [20–22].

High Dependence on Annotated Data

In 2003, scientists collaboratively initiated the Encyclopedia of DNA Elements (ENCODE) project to explore and decode a vast array of functional elements within the human genome [23]. By leveraging bioinformatics techniques, genome annotation involves identifying a wide range of functional elements, including coding genes, non-coding RNAs, repetitive sequences like transposons, and regulatory elements [24].

In early genome annotation methods, traditional approaches often use hybridization-based techniques [25,26] or experimental methods [27,28], which significantly rely on human knowledge and expertise. Different bioinformatics software tools including Blast2GO, InterProScan and GeneMark, has been applied for gene annotation [29–31]. However, these methods and software offer only limited contributions, lacking the capacity to effectively process high-throughput data. This presents a major challenge in bioinformatics, particularly in the analysis of large-scale omics data.

Reliance on annotation data is a key limitation of gene sequence and structure analysis. High-quality annotation is essential for accurate functional predictions and evolutionary inferences. However, the availability and completeness of such annotations can vary greatly between species, leading to inconsistent data interpretation. Many genome databases remain poorly annotated, especially for non-model organisms, and this deficiency can hinder research efforts. In addition, the manual annotation process is labor-intensive and prone to human error, resulting in inaccurate predictions of gene function. This dependence on curated data underscores the need for automated annotation tools and methodologies to enhance the efficiency and accuracy of genome analysis, particularly as sequencing data continues to expand at an unprecedented rate.

Low Computational Efficiency and Difficulty in Scaling to Multi-Species Genomes

Most existing computational tools necessitate an understanding of the optimal evolutionary distance for selection, along with the development of new algorithms for alignment, conservation analysis, and result visualization. Traditional alignment and analysis methods often require a lot of computing resources and time and therefore are not suitable for large-scale studies involving multiple species. Limited computational efficiency remains a major challenge in gene sequence and structure analysis, hindering the processing of large-scale genomic data, especially when dealing with large genomic datasets [32].

Historically, the main limitation in genomic analysis was the sequencing process itself, that was significantly more costly than computational analysis. However, as computational analyses continue to advance, there is a growing need to enhance sequencing throughput while reducing costs, leading to the generation of even larger volumes of genomic data, as a result, computational cost and efficiency have become increasingly critical in the analysis of [33]. The complexity of analyzing algorithms hindered the ability to analyze and interpret genomic information in a timely manner.

To comprehend the complexities of genome organization, gene function, and the genetic mechanisms underlying diseases, there is an urgent need to develop more efficient algorithms and computational frameworks. As the field shifts toward more comprehensive analyses, including multi-species comparisons, these tools must be capable of managing the scale and intricacy of modern genomic data. The rapid growth of genomic data presents major challenges in terms of computational resources and analysis methods. Traditional techniques need to be improved to handle the huge amounts of data required for detailed comparative studies. High-Performance Computing (HPC) has become a key solution to these issues, with Graphics Processing Units (GPUs) playing a central role. Their ability to process many tasks simultaneously helps to significantly speed up time-consuming computational processes. Furthermore, the advent of machine learning and high-performance computing may provide promising solutions to these challenges, allowing researchers to more effectively explore the vast potential of genomic data [33].

Progress in Gene Prediction Models Based on Deep Learning

Protein and Gene Language Models

Today, Artificial Intelligence (AI) systems primarily depend on machine learning and deep learning. Machine learning enables systems to learn from data, automating the creation of analytical models and solving complex tasks with minimal human input. Deep learning, a subset of machine learning based on artificial neural networks, has demonstrated outstanding performance across various applications, often surpassing traditional machine learning models and conventional data analysis techniques.

In gene prediction, deep learning has been particularly transformative, driving advancements in protein and genomic language models that leverage vast biological datasets to uncover intricate relationships between sequences and their functions. The combination of deep learning and modern DNA/RNA sequencing technologies has transformed the field of bioinformatics, opening new frontiers in genomics research and accelerating the translation of massive biological data into actionable insights. As these AI-driven approaches continue to evolve, they promise to unlock deeper insights into gene regulation, protein interactions, and the fundamental mechanisms underlying life at a molecular level.

Over the past few years, machine learning has led to significant advancements in the efficient analysis of preprocessing techniques. The development of artificial neural networks (ANNs) has driven the evolution of deep neural network architectures, enhancing learning capabilities and giving rise to what we now refer to as deep learning [35].

We have illustrated the classification of machine learning algorithms applied in nucleic acid structure prediction, emphasizing the transition from traditional machine learning methods to deep neural learning techniques (Figure 2). It is structured into three main levels: general machine learning algorithms, artificial neural networks (ANNs), and deep neural learning models, showcasing their hierarchical relationships and increasing complexity. Within this framework, deep learning emerges as the most sophisticated approach, utilizing deep neural networks to process genomic sequences with higher accuracy. The advanced models such as

RNA-FM and ProteinRNA-FM enhanced DNA and RNA structure prediction through deeper pattern recognition and improved accuracy. These progressive advancements, from traditional machine learning models to sophisticated deep neural networks, demonstrate the growing role of artificial intelligence in DNA and RNA structure prediction. These advancements provide a powerful toolkit for genomic research, molecular biology, and bioinformatics applications, enabling more precise and efficient structural analysis.

Since the emergence of deep learning methods, the landscape of gene prediction has been greatly changed. These models leverage vast biological data to unravel the intricate relationships between sequences and their biological functions, successfully predicting the structures of nucleic acids and proteins [31 33 34].

Here we provided a brief list of tools and algorithms for variant calling and annotation, along with links to their source code in TABLE 2. (modified from [36]) to aid in choosing the most appropriate deep learning tool for a specific data type. This table offers a summary of different deep learning models used in omics research, highlighting their specific architecture, target datasets, and predictive purposes. It includes models such as RNNs, CNNs, LSTMs, GANs, and AE-based frameworks, demonstrating their application in miRNA target prediction, gene expression analysis, histone modification classification, variant calling, and epigenetic variation detection.

Traditional methods and deep learning-based approaches offer distinct advantages and challenges when applied to various fields. Traditional methods, often based on predefined rules and handcrafted features, typically provide interpretable results but may struggle with complex patterns and large datasets. In contrast, deep learning methods excel at capturing intricate patterns and handling vast amounts of data through automatic feature extraction. When comparing these approaches using quantitative metrics, deep learning typically delivers superior accuracy by effectively capturing complex, non-linear relationships. However, this improved accuracy can come at the cost of computational efficiency, as deep learning models usually require significant processing power and longer training times. Traditional methods, on the other hand, tend to be more computationally efficient and require fewer resources, making them suitable for applications where real-time processing is crucial. Strengthening the argument with such quantitative comparisons provides a clearer perspective on selecting the appropriate approach based on specific requirements.

To make the comparison clearer, let's take the example of image classification. We can contrast a traditional machine learning approach, such as a Support Vector Machine (SVM) with handcrafted features, against a deep learning model like a Convolutional Neural Network (CNN).

Precision: Traditional method (SVM with Histogram of Oriented Gradients features): ~85% accuracy on a benchmark dataset (e.g., MNIST). Deep learning method (CNN with multiple convolutional layers): ~99% accuracy on the same dataset. Analysis: Deep learning demonstrates significantly higher accuracy, especially as the complexity of the dataset increases.

Computational Efficiency: Training time for SVM: A few minutes on a standard CPU. Training time for CNN: Several hours to days on a GPU, depending on the network depth and dataset size. Inference time for SVM: A few milliseconds per image. Inference time for CNN: A few milliseconds to seconds, depending on the model size.

Analysis: Traditional methods are more efficient in terms of computation, making them preferable for applications with limited hardware resources. By presenting such quantitative comparisons, it becomes evident that while deep learning models offer superior accuracy, their computational demands can be a limiting factor, necessitating a careful choice based on application-specific constraints.

Due to these characteristics, deep learning has become increasingly applied in the biological analysis of DNA, RNA, and proteins. Its ability to automatically extract complex features and identify intricate patterns from large-scale biological datasets, which makes it particularly well-suited for tasks such as sequence classification, motif discovery, and structural prediction. The higher accuracy of deep learning models enables more precise identification of functional elements and interactions within biological sequences, which is crucial for advancing genomics, transcriptomics, and proteomics research. Despite the computational challenges, the potential to uncover novel insights and drive discoveries has made deep learning an indispensable tool in modern biological data analysis.

For nucleic acids, there are multiple models belonging to DL model for prediction, The TABLE 3 presents a list of deep learning methodologies in genomics. From left to right, the columns detail the acronym of the deep learning (DL) model (if applicable), the DL model utilized, the omics data used as input, the prediction or research objective, and the corresponding evaluation metrics (modified from [37]).

Other BERT based models like DNABERT [38], CDSBERT [39] and ProtBERT (adapted for nucleotide prediction by retraining on DNA/RNA sequences) [40]. Thess models encompasses both local and global representations, enabling end-to-end processing of these inputs and outputs, and allowing the prediction of their functions based on learned contextual relationships.

This approach allows researchers to decode complex patterns within genomic data, facilitating the identification of potential pathogenic genes and their roles in diseases. Furthermore, the integration of these models Like DNABERT-2 [41], a pre-trained BERT model learns representations of DNA k-mers by treating sequences as a language and incorporating both upstream and downstream nucleotide contexts, genomic studies have achieved more accurate and efficient predictions, thereby improving our understanding of gene regulation and expression [42]. BERT-based models are proving transformative in nucleotide prediction, enabling breakthroughs in genomics, regulatory analysis, and functional annotation.

Here are brief definitions and contexts for some advanced deep learning models used in bioinformatics and genomics:

CDSBERT (Coding DNA Sequence BERT): A transformer-based model specifically pre-trained on coding DNA sequences to understand sequence patterns, codon usage, and functional elements. It is developed to enhance tasks like gene prediction, variant effect analysis, and gene function annotation. DNABERT (DNA Bidirectional Encoder Representations from Transformers): A model of BERT-based, which pre-trained on genomic sequences using k-mers tokenization, which allows it to capture DNA sequence context more effectively than traditional one-hot encoding. DNABERT is widely used for tasks such as promoter identification, mutation classification, and sequence alignment. ProtBERT (Protein BERT): A transformer model trained in protein sequences, capturing amino acid-level dependencies to enhance protein function prediction, structure classification, and evolutionary analysis. RNA-BERT: A specialized BERT model pre-trained on RNA sequences, focusing on secondary structure and regulatory elements crucial for understanding post-transcriptional modifications and RNA-protein interactions.

A key challenge in applying deep learning to biological analysis is the quality and scarcity of available data. Biological data, including DNA, RNA, and protein sequences, often suffer from noise, missing values, and inconsistencies due to limitations in experimental techniques and variability in sample conditions. Moreover, acquiring high-quality labeled biological data can be both costly and time-intensive to obtain, often requiring expert curation and validation. Considering the complexity of biological systems, the available datasets are often relatively small, leading to challenges such as overfitting and limited generalizability on conventional machine learning models.

To address these challenges, deep learning models, particularly protein and gene language models, have become powerful tools. Inspired by techniques of natural language processing, these models treat biological sequences as "languages," capturing intricate dependencies and patterns within genomic and proteomic data. Trained on large volumes of unlabeled biological sequences, these models can effectively learn meaningful representations and be fine-tuned for specific tasks, such as protein structure prediction, gene function annotation, and variant effect analysis. Additionally, techniques like transfer learning and data augmentation enable these models to generalize better even with limited labeled datasets, offering significant advantages in extracting biological insights. Despite their potential, careful model validation and integration with experimental data remain essential to ensure the reliability and interpretability of deep learning-driven biological discoveries.

Applicability of the BERT Architecture in Gene Language Modeling

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model for contextualized language representation, which has achieved performance surpassing human-level accuracy in many natural language processing (NLP) tasks. It introduces a pre-training and fine-tuning approach, where the model first learns a general understanding from vast amounts of unlabeled data and then leverages task-specific data to address a wide range of applications with minimal changes to its architecture [43].

In 2017, A. Vaswani introduced BERT architecture, which is based on a multilayer bidirectional transformer [44]. This architecture trained two versions of the neural network: a standard version with 12 layers and 768 dimensions (containing 110 million parameters in total) and a larger version with 24 layers and 1024 dimensions (containing 340 million parameters). BERT uses text embeddings to represent an input sequence, where a sequence is defined as an arbitrary set of contiguous text tokens.

Developed by Google and released in 2018, BERT is designed to understand the context and nuances of human language in ways never seen before [45]. By utilizing a bidirectional approach, BERT can understand the nuances of biological sequences, such as the dependencies between nucleotides or amino acids, which are crucial for accurate gene prediction. One of the standout features of BERT is its pretraining process. It is trained on a vast corpus of text from the internet, enabling it to absorb an extensive amount of knowledge. This pre-trained model can then be fine-tuned for a wide range of natural language understanding tasks, such as text classification, question answering, and language translation [46]. Belong to Large Language Models (LLMs), this architecture empowers the model to revolutionize the field by tackling challenges related to large, complex biological datasets, enabling it to make informed predictions about gene functions and interactions [47]. As a result, BERT-based models together with other language models, have become a cornerstone in the field of computational biology, driving advancements in gene prediction and protein prediction methodologies.

Development and Application of ProtBERT, CDSBERT, and DNABERT

DNABERT, CDSBERT, and ProtBERT represent significant advancements in the application of deep learning to biological sequences.

DNABERT, on the other hand, is tailored for DNA sequences, allowing for the analysis of genomic data with unprecedented accuracy. When compared to the most used genome-wide regulatory element prediction programs, DNABERT has shown superior ease of use, accuracy, and efficiency. Experimental results demonstrate that a single pre-trained Transformer model can simultaneously achieve state-of-the-art performance in predicting promoters, splice sites, and transcription factor binding sites. Furthermore, DNABERT can directly visualize nucleotide-level importance and semantic relationships within the input sequence, enabling improved interpretation and accurate identification of conserved sequence motifs and functional genetic variations. DNABERT is not only applied to humans, but also to many other organisms through fine-tuning and has excellent performance [38,48].

CDSBERT focuses on coding sequences, enhancing the prediction of gene functions by integrating coding sequence data into its training process. CDSBERT variants created a highly biochemically relevant latent space, surpassing their amino acid-based counterparts in predicting enzyme commission numbers. Further analysis showed that synonymous codon token embeddings shifted noticeably in the embedding space, highlighting the distinct information added across a broad phylogeny within these traditionally considered "silent" mutations [39].

ProteinBERT is specifically designed for protein sequences, employing a transformer architecture to capture the contextual information of amino acids. ProteinBERT, a deep language model tailored for proteins, integrates language modeling with the novel task of Gene Ontology (GO) annotation prediction. Its architecture incorporates both local and global representations, enabling end-to-end processing of these inputs and outputs. ProteinBERT achieves multiple benchmarks across a range of protein properties, including protein structure, post-translational modifications, and biophysical characteristics [40]. AggBERT achieved state-of-the-art performance, highlighting the potential of large language models to enhance the accuracy and speed of Amyloid Fibril prediction, surpassing traditional heuristics and structure-based methods [49]. Overall, based on ProteinBERT, it offers an efficient framework for quickly training protein predictors, even when there is limited labeled data available.

These models have been successfully applied to various genomic tasks, including predicting gene functions and identifying regulatory elements, showcasing their versatility and robustness in handling diverse biological data. The development of these models marks a pivotal step towards automating and improving the accuracy of gene prediction, thereby accelerating research in genomics and molecular biology.

Integration of Diverse Features

Comprehensive analysis of sequence, structure and function information

Multi-view data often offers unprecedented opportunities to gain insights into complex biological systems from multiple perspectives and levels. However, it presents a significant challenge for data experts and scientists to effectively optimize these datasets for specific needs. Integration, which means combining various features, including sequence, structure, and function information, is essential to enhance the performance of gene prediction models that integrate sequence information, transcription factor binding, histone modifications, chromatin accessibility, and 3D genome data, providing a comprehensive framework for more efficient subsequent analysis [50].

In early integration methods, features from various data sources are merged into a unified feature vector. In contrast, late integration involves training separate models for each data view and then combining their outputs to make the final decision [51]. By integrating different types of data, researchers can develop more comprehensive models that account for the multifaceted nature of gene regulation and expression. Integrating structural data enables models to account for the three-dimensional conformation of proteins, which is crucial for comprehending their functions [52]. In addition, incorporating functional annotations and evolutionary data can further refine predictions, allowing the model to leverage historical biological information to improve accuracy. This comprehensive approach not only boosts the model's predictive accuracy but also provides a deeper insight into the biological significance of the predicted genes.

Models Combining RNA Secondary and Tertiary Structure Information

RNA structure plays a crucial role in various processes, including ligand sensing, as well as the regulation of translation, polyadenylation, and splicing. The mRNA structures of genes involved in cellular functions and stress responses often possess features that enable these RNAs to undergo conformational changes in response to environmental factors [53]. RNA structure is pivotal in the post-transcriptional regulation of gene expression, influencing processes such as RNA maturation, degradation, and translation. With the advent of next-generation sequencing, RNA structure research has evolved from in vitro low-throughput probing methods to in vivo high-throughput RNA structure profiling. The advancement of these techniques facilitates ongoing studies into the functional roles of RNA structure [54].

Moreover, in recent years more advancements have also focused on models that incorporate RNA secondary and tertiary structure information, such as RNA-FM [55], PlantRNA-FM [56,57], and PINC [58]. These models use structural data to improve predictions of RNA function and interactions, acknowledging that the spatial configuration of RNA molecules is essential for their biological functions. By combining secondary structural features with sequence data, more accurate predictions of RNA behavior can be made. This integration is particularly important given the complexity of RNA structures, which often involve intricate folding patterns that are essential for their functionality. As such, the development of models that incorporate both sequence and structural information represents a pivotal advancement in

RNA biology [55].

Different Basement Combining RNA Secondary and Tertiary Structure Information

In order to gain a deeper understanding of the structure of RNAs, the RNA-FM and PlantRNA-FM were performed based on Bert and Transformer, respectively. The original Transformer model can be traced back to 2017 [59]. The core innovation of the Transformer model lies in its introduction of self-attention mechanisms. A key challenge in RNA secondary structure prediction is accurately identifying interactions between nucleotides that form stem-loop structures. Transformers address this challenge by enabling each nucleotide to attend to every other nucleotide in the sequence, and this capability makes the Transformer particularly well-suited for modeling such interactions. This is mathematically defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In RNA sequence prediction, Q, K, and V represent query, key, and value vectors from nucleotide embeddings, identifying nucleotides of interest, interaction partners, and encoded information, respectively. The SoftMax function ensures attention scores sum to 1, enabling predictions of base-pairing probabilities and structural motifs like stems, bulges, and loops. A practical enhancement incorporates sequence-specific constraints into the attention mechanism by modifying the score calculation using a constraint matrix C, which prioritizes biologically plausible interactions.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + C}{\sqrt{d_k}}\right)V$$

To enhance the model's ability to capture diverse data patterns, Transformers employ multi-head attention, which computes multiple attention functions in parallel and combines their outputs:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each head is calculated as: $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

The Transformer model uses positional encoding to embed token position information into the sequence. This encoding is added to the input embeddings and computed as follows:

$$\begin{aligned} PE(pos,2i) &= \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right), \\ PE(pos,2i+1) &= \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \end{aligned}$$

BERT (Bidirectional Encoder Representations from Transformers) is based on Transformer proposed by Google in 2018， the emergence of BERT marks an important milestone in pre-trained

language models [60]. BERT addresses this limitation with a bidirectional encoder, enabling it to capture both upstream and downstream dependencies simultaneously. This is accomplished through Masked Language Modeling (MLM), a pretraining task in which random input tokens are masked, and the model learns to predict them based on the surrounding context:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{x\sim\mathcal{X}}\mathbb{E}_{\mathcal{M}\sim x}\sum_{i\in\mathcal{M}} -\log p(x_i \mid x/\mathcal{M})$$

M represents the set of masked tokens in the input sequence x, and p(xi|x/M) denotes probability of predicting the original token xi given the masked input. In addition to MLM, BERT also incorporates a Next Sentence Prediction (NSP) task, which further strengthens its ability to understand sentence relationships. In this task, the model is provided with two sentences and must determine whether the second sentence logically follows the first. The loss function for NSP is defined as:

$$\mathcal{L}_{\text{NSP}} = -\left[y\log P(\text{IsNext}) + (1-y)\log P(\text{NotNext})\right]$$

Where y is a binary label indicating whether the second sentence is the next sentence or a randomly sampled one, and P (IsNext) and P (NotNext) are the model's predicted probabilities.

Comparation and limitations Transformer and BERT

The Transformer and BERT are pivotal architectures in natural language processing, differing in design and applications. Transformers, a general architecture based on attention mechanisms, capture global dependencies in data through self-attention and multi-head attention mechanisms. They are composed of an encoder that processes input sequences and a decoder responsible for generating the output. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model based on the Transformer encoder. Its key innovation is bidirectional encoding, enabling it to consider contextual information from both directions during training, excelling in sentence structure and semantic understanding (TABLE 4. indicate the difference). In terms of application scenarios Transformer, due to its flexible architecture, BERT is widely used for tasks like machine translation and text generation, whereas BERT is primarily applied in natural language understanding tasks such as text classification and named entity recognition. Large Language Models (LLMs) have revolutionized natural language processing (NLP), setting new benchmarks across a wide range of tasks. Transformer-based models like BERT and GPT employ pooling layers to combine token-level embeddings into more comprehensive sentence-level representations, which are essential to their overall performance [61]. Missing data is a common challenge in wireless networks and many other fields, often limiting the effectiveness of machine learning and deep learning models. FGATT tackles this problem by combining the Fuzzy Graph Attention Network (FGAT) with the Transformer encoder, offering a strong and accurate solution for data imputation [62].

In RNA structure prediction, the difference between Transformer and BERT lies in their handling of long-range dependencies and contextual understanding. Transformers excel at capturing distant nucleotide interactions, such as base pairing, by employing self-attention, which allows each token to focus on all other tokens simultaneously. BERT’s bidirectional encoding makes BERT particularly effective for tasks such as predicting RNA-binding sites and analyzing mutation effects on RNA structures, making it well-suited for RNA sequence inference tasks.

In summary, Transformers are celebrated for their flexibility and strong generative capabilities, while BERT stands out in natural language understanding tasks. Their complementary strengths open new possibilities and drive innovations in natural language processing [63,] Self-GenomeNet. In different cases of analyzing, the basement model should be selected flexibly.

TABLE 4. compares Transformer and BERT architectures, highlighting their differences in structure and application. Transformers use an encoder-decoder design for generative tasks like translation, while BERT is encoder-only, excelling in text understanding. Both employ self-attention; however, BERT employs masked language modeling (MLM) and next sentence prediction (NSP) during pretraining. These distinctions make Transformers better for sequence generation and BERT for contextual language comprehension.

Significant Improvements for Model Performance in application

Deep learning has made incredible progress in recent years. It works by combining many complex functions to understand the relationships between input data and results. Although neural networks have been around for a long time, recent improvements have greatly boosted their performance in areas like computer vision and natural language processing. The optimization of deep learning models has also led to better accuracy, greater ability to apply to different situations, and faster processing.

For example, Deep Dual Enhancer is a method based on DNABert that uses a combination of multi-scale convolutional neural networks and BiLSTM to identify enhancers. The optimization of DNABERT-2 has resulted in enhanced performance metrics, including accuracy and speed, allowing for more efficient processing of large genomic datasets [65]. These improvements are critical for practical applications. The ability to quickly and accurately predict gene function will have a profound impact on fields such as personalized medicine and genetic engineering. In addition, advances in computational efficiency allow researchers to analyze larger data sets, facilitating the exploration of complex biological problems that were previously unattainable. As deep learning continues to develop and be applied, there is still great potential to further improve model performance.

In summary, the combination of deep learning techniques and genetic prediction models has revolutionized the field, offering new insights and capabilities that deepen our understanding of genetic information and its role in health and disease. As these models continue to be developed and applied, they will continue to shape the future of genomic research.

Innovations in Datasets and Modeling Techniques

Data Preprocessing

Data preprocessing is an essential step when analyzing genomic and transcriptomic data because it directly affects the quality and accuracy of the results. In genomic research, preparing the data is especially important when using machine learning methods for genomic selection. This process involves cleaning the raw data, adjusting it to a consistent scale, and transforming it to make sure it's ready for analysis. For example, data preprocessing for next-generation bisulfite sequencing involves data cleaning, normalization, and addressing batch effects. Additionally, reduced representation bisulfite sequencing requires special handling to account for adapter contamination and artificial bases incorporated into sequence reads [66]. While in transcriptomic research such as RNA-Seq, data preprocessing methods can be applied to minimize systematic variations and harmonize the datasets before they are used to build a machine learning model for tissue-of-origin classification [67].

Data preprocessing can include the removal of low-quality reads, trimming of adapter sequences, and normalization of expression levels. Advanced preprocessing techniques, such as through the application of machine learning and deep learning techniques, can further enhance data quality by identifying and correcting systematic biases in the data [37,68].

For example, the implementation of tools like LongQC provides automated quality control for long-read sequencing data, ensuring that the datasets used for analysis are of high quality and accurately represent the underlying biological phenomena quality control tool for genomic datasets produced by third-generation sequencing (TGS) technologies, such as Oxford Nanopore Technologies (ONT) and SMRT sequencing from Pacific Biosciences (PacBio) [69]. Additionally, the integration of multiple sequencing technologies, such as combining short-read and long-read sequencing, can improve the comprehensiveness of genomic datasets, allowing for more accurate variant calling and functional annotations. TABLE 4. indicates a list of deep learning methods in transcriptomics.

Annotation and Organization of Genomic and Transcriptomic Data

An organism's genome impacts nearly every facet of human biology, from molecular and cellular functions to health and disease phenotypes. Investigating DNA sequence variations between individuals (genomic variation) can uncover previously unknown biological mechanisms, identify genetic factors contributing to disease susceptibility, and aid in the development of new diagnostic tools and therapies [70]. The annotation and organization of genomic and transcriptomic data are essential for comprehending the functional implications of genetic variation. Accurate annotation involves linking genomic sequences to their corresponding biological functions through various computational tools and databases.

For example, the use of Human Phenotype Ontology (HPO) enables researchers to standardize clinical features and predict causative genes from phenotypic data; the interpretation of genomic variants from whole exome sequencing (WES) can be improved by using Human Phenotype Ontology (HPO) terms, which help standardize clinical features and predict causative genes [71]. Furthermore, tools like Phen2Gene leverage a database based on this information, known as the HPO2Gene Knowledgebase (H2GKB). This resource provides weighted and ranked gene lists for each HPO term, aiding in the identification of potential pathogenic variants in genomic studies [72]. The integration of these annotation tools with machine learning approaches enhances the ability to predict phenotypic outcomes from genomic data, thereby enhancing the understanding of genotype-phenotype relationships. The empirical Bayes prior, which is expected to align with the observed epistasis pattern, is used to reconstruct the genotype–phenotype mapping through Gaussian process regression [73,74].

Data Compression Techniques

(e.g., Byte Pair Encoding as a Replacement for k-mer Markers)

Large biological datasets are being generated at an accelerating rate, posing significant storage challenges, especially in the field of high-throughput sequencing (HTS) [75]. Data compression technology is crucial for managing the vast amounts of genomic data generated by high-throughput sequencing technologies. Efficient compression not only reduces storage requirements but also enhances data transmission speed, which is crucial for large-scale genome research.

One approach to storing and indexing datasets is by using sets of k-length substrings, known as k-mers [76]. BPE can effectively compress genomic sequences by identifying and replacing frequently occurring pairs of bytes with shorter representations, thereby reducing the overall size of the dataset without losing critical information by [77]. RBFQC outperforms other state-of-the-art genome compression methods. When compared to GZIP, RBFQC achieves a compression ratio of 80-140% for fixed-length datasets and 80-125% for variable-length datasets. In comparison to domain-specific FastQ file referential genome compression techniques, RBFQC offers a 10-25% improvement in both compression and decompression speeds [78].

All the methods have been used to outperform conventional compression techniques, achieving significant reductions in file size while maintaining the integrity of the genomic data. Such advancements in data compression are essential for facilitating the analysis and sharing of genomic information across research platforms.

Model Optimization

Model optimization is a key aspect of enhancing the predictive power of genomic and transcriptomic analyses. Various strategies can be employed to optimize models, including hyperparameter tuning, feature selection, and the integration of ensemble methods. Hyperparameter tuning is a critical step in finding the optimal machine learning parameters. Identifying the best hyperparameters can be time-consuming, especially when the objective functions are expensive to compute or when many parameters need to be adjusted [79].

Bayesian optimization is a technique used to optimize objective functions that are time-consuming to evaluate, often requiring minutes or hours. It is particularly well-suited for optimizing continuous domains with fewer than 20 dimensions and can handle stochastic noise in function evaluations. The method constructs a surrogate model for the objective and quantifies the uncertainty of this model using a Bayesian machine learning technique called Gaussian process regression. An acquisition function, derived from this surrogate, is then used to determine the next step [80]. As an alternative to convolutional neural networks (CNNs), vision transformers (ViT) offer strong representational power through spatial self-attention and channel-level feed-forward networks. The ability to model relationships across spatial and channel dimensions sets ViT apart from other networks. To enhance this capability, we introduce Feature Self-Relation (SERE) for training self-supervised ViT models, resulting in stronger representations that consistently improve performance on a range of downstream tasks [81]. Additionally, the incorporation of epistatic interactions in models, such as Epistatic Random Regression BLUP (ERRBLUP), which accounts for all pairwise SNP interactions, and selective Epistatic Random Regression BLUP (sERRBLUP), which focuses on a selected subset of pairwise SNP complex interactions, further enhances the model's capability [82]. GBLUP, ERRBLUP, and sERRBLUP are applied using genotypes from a publicly available wheat dataset along with their respective simulated phenotypes. These optimization techniques play a crucial role in developing robust models that can accurately forecast phenotypes based on genomic data [83].

Application of Self-Supervised Learning Methods

Self-supervised learning methods have become a powerful tool for extracting meaningful insights from genomic data, all without requiring large, labeled datasets. By leveraging the inherent structure of the data, self-supervised learning (SSL) can enhance the performance of predictive models, particularly when labeled data is limited. SSL has become a powerful technique for extracting meaningful representations from vast, unlabeled datasets, driving advancements in computer vision and natural language processing. It holds great potential for applications in single-cell genomics (SCG), enabling fully connected networks and benchmarking their utility across key representation learning scenarios [84].

By leveraging unlabeled data, self-supervised learning techniques can enhance the performance of machine learning models, especially when labeled data is limited. Self-GenomeNet, a self-supervised learning method specifically designed for genomic data, is ideal for large-scale, unlabeled genomic datasets and has the potential to significantly improve the performance of genomic models. Self-GenomeNet utilizes reverse-complement sequences to learn dependencies and improve prediction accuracy in data-scarce genomic tasks for improvement. This approach not only improves the efficiency of model training but also aids in the discovery of novel genetic associations and biomarkers, ultimately advancing our understanding of complex biological processes. This new approach not only enhances the efficiency of training models but also facilitates the discovery of novel genetic associations and biomarkers, ultimately enhancing our comprehension of intricate biological processes [85].

Optimized Model Architectures for Long Sequence Processing
(e.g., Attention with Linear Biases)

The development of optimized model architectures for processing long sequences is essential for advancing genomic and transcriptomic analyses. For instance, machine learning is widely used in genomics to identify patterns in data and generate new biological hypotheses [86]. Traditional models often struggle with the inherent challenges posed by long sequences, such as increased computational complexity and memory requirements.

Since Vaswani introduced the transformer model [44], It was initially demonstrated that extrapolation could be achieved by simply altering the position representation method. However, the current methods still do not support efficient extrapolation. A simpler and more efficient approach called Attention with Linear Biases (ALiBi) was introduced. This method biases the query-key attention scores by applying a penalty proportional to their distance. ALiBi's inductive bias favoring recency allows it to outperform several powerful positional methods on the WikiText-103 benchmark. These methods were specifically designed to tackle challenges by enhancing the efficiency of attention mechanisms in deep learning models [87]. This architecture allows for the effective processing of long genomic sequences while maintaining high predictive accuracy.

In conclusion, by optimizing model architectures in this way, researchers can enhance their ability to analyze complex genomic data and extract meaningful biological insights with improved efficiency.

Cross-Species Genomic Analysis

Cross-species genomic analysis is an effective method for understanding evolutionary relationships and functional conservation across different organisms. By combining genomic data from multiple species with comparative genomics techniques, researchers can identify conserved genetic elements and pathways that play crucial roles in various biological processes [88].

Recent studies have demonstrated the effectiveness of multi-species genomic modeling in predicting phenotypic outcomes, as seen in the analysis of hesperidia infection risk across diverse bird populations [89]. Additionally, the use of genomic data from model organisms to inform studies in non-model species facilitates the transfer of knowledge and enhances the understanding of complex traits and diseases [90]. This inter-species approach not only broadens the scope of genomic research but also contributes to the conservation of biodiversity and the study of evolutionary dynamics.

Applications of Gene Structure Prediction: Case Studies

Annotation of Functional Regions

Gene structure prediction is essential for understanding the functional elements within the genomes of different organisms. Identifying and annotating functional regions, such as promoters, enhancers, and other regulatory elements, is crucial for unraveling gene function and regulation.

Based on the BERT-based model, DNABERT-2 employs deep learning techniques to analyze DNA sequences, allowing for precise identification of functional regions like promoters and enhancers in human genes, BERT-TFBS is an innovative BERT-based model designed for predicting transcription factor binding sites through transfer learning. The model integrates a pre-trained BERT module (DNABERT-2) [91]. The model consists of a convolutional neural network (CNN) module, a convolutional block attention module (CBAM), and an output module. The BERT-TFBS model leverages the pre-trained DNABERT-2 module to capture complex long-term dependencies in DNA sequences through a transfer learning approach, while the CNN module and CBAM work together to extract high-level local features [92].

Large language models, commonly used in natural language processing like Google's BERT and OpenAI's GPT-X, have primarily been applied in fields such as genomics, transcriptomics, proteomics, drug discovery, and single-cell analysis [46].

Animals and humans: Using RNA-FM to Predict Functional Regions in Mammals

Predicting RNA secondary structure presents a major challenge for RNA structural biologists, necessitating dedicated efforts to refine our understanding of RNA folding principles and improve the accuracy of structure prediction models. These models have significant potential to advance downstream applications, such as the development of RNA-targeting drugs [93]. In the realm of animal studies, RNA-FM and RNA-MSM and related tools have become powerful methods for predicting functional regions within mammalian RNA sequences.

The pre-trained model can be further refined for a variety of downstream tasks related to RNA structure and function. The method leverages the structural properties of RNA to identify regions that are RNA 2D/3D structure prediction. RNA-FM uses self-supervised learning to predict secondary and 3D structures by leveraging the extensive dataset of non-coding RNA sequences. This approach enables RNA-FM to capture a wide range of structural information, offering a comprehensive understanding of RNA sequence features. In contrast, RNA-MSM utilizes homologous sequences obtained from RNAcmap through an automated pipeline. This model excels at accurately mapping to 2D base pairing probabilities and 1D solvent accessibility [46].

By analyzing the RNA secondary structures, RNA-FM uses self-supervised learning to predict secondary and 3D structures by leveraging the extensive dataset of non-coding RNA sequences. This approach enables RNA-FM to capture a wide range of structural information, offering a comprehensive understanding of RNA sequence features. In contrast, RNA-MSM utilizes homologous sequences obtained from RNAcmap through an automated pipeline. This model excels at accurately mapping to 2D base pairing probabilities and 1D solvent accessibility, predicting RNA splicing in the SARS-CoV-2 genome structure and evolution, modeling protein-RNA binding preferences, and modeling gene expression regulation. RNA-FM can predict how these structures influence protein binding sites, thereby enhancing our understanding of post-transcriptional regulation in mammals [55].

Furthermore, the study of CodonBERT defined a BERT-based architecture for codon optimization using the cross-attention mechanism, presenting a model specifically developed for codon optimization, an essential component in the design of mRNA vaccines. This research highlights the potential of BERT-based architectures in optimizing codon sequences to enhance protein expression, which is directly relevant to vaccine development. While these studies do not directly apply RNA-BERT to vaccine design, they demonstrate the adaptability of BERT-based models in RNA sequence analysis, suggesting potential applications in RNA vaccine development [94].

This predictive capability is particularly valuable for studying complex gene regulatory networks and understanding how alterations in RNA structure can impact gene expression and contribute to various diseases.

Plants: Analyzing the Role of PlantRNA-FM in RNA Secondary Structures

In animals, this model is well established for RNA structure prediction. Studies conducted in Arabidopsis thaliana have shown that specific RNA structures can influence the transcriptional activity of genes, thus, it plays a crucial role in plant development and its response to environmental stimuli. In plant biology, the double-stranded RNA structures downstream of uAUGs (referred to as uAUG-ds) play a key role in the selective translation of uAUGs, enabling the prediction and rational design of translating uAUG-ds. The widespread use of deep learning-based RNA structural features and the conservation of RNA helicases across different kingdoms suggest that mRNA structural [95].

Then PlantRNA-FM has been utilized to investigate the relationship between RNA secondary structures and transcriptional regulation. It was pretrained on a vast dataset, incorporating RNA sequences and structural data from 1,124 distinct plant species. PlantRNA-FM demonstrates exceptional performance in plant-specific downstream tasks. This model enables the exploration of functional RNA motifs within the complex plant transcriptomes, providing plant scientists with the tools to program RNA codes in plants. The application of PlantRNA-FM allows researchers to predict how RNA secondary structures can affect gene expression and regulatory mechanisms, providing insights into plant adaptation and evolution [56].

Humans: Precise Prediction of Human Gene Functional Regions with DNABERT-2

DNABERT-2 has proven to be an asset in the precise prediction of human gene functional regions, including promoters and enhancers. By utilizing a deep learning approach, DNABERT-2 can analyze vast genomic datasets to identify critical regulatory elements that govern gene expression.

Integrating DNA breathing features with the DNABERT-2 foundational model significantly improved the accuracy of TF-binding predictions. A study by researchers at Los Alamos National Laboratory integrated DNA breathing features with the DNABERT-2 foundational model enhances the prediction of transcription factor binding. This approach enhanced the accuracy of predicting gene-binding locations, offering valuable insights into mutations associated with cancer. Since gene expression can be linked to diseases like cancer, predicting the transcription factors that bind to specific gene locations could have significant implications for drug development [96]. Building on the Extended Peyrard-Bishop-Dauxois (EPBD) nonlinear DNA dynamics model, the multi-modal deep learning model EPBDxDNABERT-2 significantly enhances the prediction of over 660 TF-DNA interactions. This improvement results in an increase of up to 9.6% in the area under the receiver operating characteristic (AUROC) metric, compared to the baseline model. EPBDxDNABERT-2 could enhance predictive accuracy for disease-related non-coding variants identified in genome-wide association studies. This model integrates various genomic features and enables researchers to uncover the complexities of gene regulation in humans, opening new avenues for progress in personalized medicine and targeted therapies [97].

Protein-RNA Interactions by Deep Learning

Protein–RNA interactions are crucial for numerous cellular processes and studying them is essential for understanding the molecular mechanisms of gene regulation. Recent advancements in computational methods have improved the prediction of RNA-binding sites, enabling the identification of protein-RNA interaction networks linked to various diseases.

Recently, AlphaFold has brought a groundbreaking transformation into the field of protein biology. Looking ahead, the prediction of protein-RNA interactions is expected to see significant advancements in the coming years. This includes improvements in predicting both binding sites and binding preferences, as well as a comprehensive exploration of commonly used datasets, features, and models. The incorporation of deep learning models has greatly enhanced the accuracy of RNA-protein binding site predictions, enabling researchers to pinpoint potential therapeutic targets and gain a deeper understanding of the roles specific proteins play in RNA metabolism and gene expression regulation [98]. Advancements in protein-RNA interaction prediction have been significantly driven by deep learning technologies, enabling a deeper understanding of RNA-binding proteins (RBPs) and their regulatory roles across multiple species. Just as comprehensive clinical studies on ectopic pituitary adenomas have improved the diagnosis and management of complex diseases [99], deep learning-powered models such as RNA-FM and ProteinRNA-FM have revolutionized RNA-protein binding site predictions, enhancing accuracy in identifying functional RNA regions. Similarly, long-term clinical data accumulation, as shown in the 13-year study on pediatric pituitary neuroendocrine tumors, underscores the importance of large-scale datasets for improving model generalization [100]. Deep learning models trained on multi-species RNA-protein interaction data have demonstrated the ability to generalize across diverse biological systems, leading to more comprehensive functional annotations and a deeper understanding of post-transcriptional gene regulation.

Genomic Evolution and Diversity Studies

An analysis of genomic evolution and diversity is essential for understanding the evolutionary trends among species. DNABERT has been utilized to analyze genomic sequences across multiple species, allowing researchers to study the variations and evolutionary patterns that exist within gene structures.

For example, research has demonstrated that genomic sequences display considerable structural variations, offering valuable insights into the evolutionary history of various species and multiple ecotypes [101]. By employing advanced computational methods, researchers can effectively analyze these variations, contributing to our understanding of evolutionary biology and the mechanisms driving genomic diversity [67].

A specific case study utilizing BERTPhylo involved analyzing genomic sequences across various species to investigate evolutionary trends. Built on the established PlantSeqs dataset, with a focus on Embryophyta, this new software has demonstrated for the first time that phylogenetic trees can be constructed by automatically selecting the most informative regions of sequences, eliminating the need for manual selection of molecular markers. This finding provides a solid foundation for further exploration into the functional roles of different regions of DNA sequences, deepening our understanding of biology. This model was employed to identify structural changes in genomic sequences, revealing patterns of conservation and divergence that are critical for understanding the evolutionary relationships among species. This analysis not only expands our understanding of genomic evolution but also establishes a framework for future research focused on uncovering the genetic foundations of adaptation and speciation [102].

In conclusion, the applications of gene structure prediction, particularly through advanced computational models like DNABERT, RNA-FM, and its homologous BERT-derived algorithms, have significantly deepened our insight into gene function, regulation, and evolutionary dynamics across different organisms. These tools offer valuable insights into the complexities of genomic architecture and its impact on health and disease, laying the foundation for future research and therapeutic developments.

Prospects of Gene Structure Prediction in Disease Research and Treatment

Potential for Curing Human Diseases

Gene and protein structure prediction holds immense potential for revolutionizing the treatment and understanding of human diseases. By accurately predicting gene structure, researchers can identify disease-associated mutations and their functional implications, paving the way for new therapeutic strategies [103,104]. For example, determining the structure or type of RNA is crucial for RNA-based therapeutics, such as mRNA vaccines, RNA interference, and CRISPR-based therapies. Traditionally, RNA's three-dimensional (3D) structures have been assessed through experimental methods, including nuclear magnetic resonance, X-ray crystallography, and cryogenic electron microscopy. However, these approaches are costly and time-consuming [55,105]. As a result, computational approaches are developed and applied work as a bridge for the gap. Advances in computational methods enable the integration of multi-omics data, thereby deepening our understanding of complex diseases such as cancer and neurodegenerative disorders. The ability to predict how genetic variations affect protein function and interactions can lead to targeted therapies that are more effective and personalized. Moreover, in recent years, deep learning approaches have surpassed traditional prediction methods, such as transcript-wise screening and principal component analysis-based predictions.

Complex human diseases, including cancers, cardiovascular diseases, and respiratory disorders, pose significant public health challenges, with environmental factors playing a crucial role in their development [106]. Genomic and molecular factors related to genes, such as genotype, mRNA expression, DNA methylation, microRNA expression, genotyping, and next-generation whole genome sequencing, have significantly advanced the study of the relationship between genomic factors and complex human diseases. These advancements enable researchers to identify disease-associated factors without bias. Beyond uncovering the molecular mechanisms behind these diseases, researchers hope that understanding these genomic factors will aid in disease diagnosis and the development of personalized treatments and new medicines [107].

Furthermore, as people develop a more thoroughly understanding of the genetic underpinnings of diseases, the potential for gene therapy and genome editing technologies, such as deep learning's impact on genomic exploration, deep learning guide CRISPR, deep learning guide RNA and protein structures [108–110]. These technologies can directly correct genetic defects, predict the target of the drugs, significantly enhancing the identification of potential drug candidates and offering the potential to cure previously untreatable conditions.

Early Diagnosis: Predicting Cancer-Related Mutations through RNA Secondary Structure

Early cancer diagnosis is crucial for improving patient outcomes, and gene structure prediction can greatly improve our ability to identify cancer-related mutations. Recent studies have indicated that RNA secondary structure plays a pivotal role in the stability and function of RNA molecules, with mutations leading to structural abnormalities that can contribute to diseases such as tumorigenesis. By employing computational tools to predict RNA secondary structures, researchers can identify mutations that disrupt these structures, offering valuable insights into the molecular mechanisms underlying cancer development. This approach not only supports early cancer detection but also promotes the development of RNA-based therapeutics and diagnostics, enabling a more proactive approach to managing diseases such as neurological disorders and cancer.

Personalized Medicine:
Predicting Drug Response Based on Patient-Specific Gene Mutations

The emergence of personalized medicine has revolutionized healthcare, enabling treatments tailored to an individual's genetic profile. In this paradigm, gene structure prediction plays a pivotal role by facilitating the identification of patient-specific mutations that impact drug response. By utilizing genomic language models, researchers can predict how these mutations affect pharmacokinetics and pharmacodynamics, which will lead to more effective and safer treatment strategies.

Building on the rapid advancements in Artificial Intelligence (AI) algorithms, recent studies have utilized domain adaptation (DA) techniques. This personalized approach not only enhances therapeutic efficacy but also minimizes adverse drug reactions, ultimately improving patient adherence and outcomes. Several learning technologies have improved the process of transferring knowledge from preclinical models to patient tumors, enabling the prediction of drugs that are specific to tumor types. These drugs exhibit higher sensitivity in tumors compared to normal tissue and show differential sensitivity across breast cancer subtypes. Furthermore, these predictions could be valuable for preclinical drug testing and phase I clinical trial design. This personalized approach not only enhances therapeutic efficacy but also minimizes adverse drug reactions, ultimately improving patient adherence and outcomes [111]. As pharmacogenomics continues to advance, integrating gene structure prediction into clinical practice will be crucial for optimizing personalized treatment strategies.

Therapeutic Tools: RNA-FM for Vaccine Design and CRISPR Optimization

The development of therapeutic tools leveraging gene structure prediction is rapidly advancing, particularly in the fields of vaccine design and gene editing.

The success of mRNA vaccines can be attributed, in part, to advancements in manufacturing that allowed to produce billions of doses with high quality and safety standards. RNA-FM supports the design of RNA vaccines by forecasting the secondary and tertiary structures of RNA molecules, which is important for eliciting robust immune responses. This methodology has been particularly relevant in the context of rapid vaccine development for infectious diseases, such as COVID-19.

The CRISPR system has driven increased investment and contributed to the discovery of additional CRISPR systems. Additionally, optimizing CRISPR technology through gene structure prediction enhances the precision of genome editing, allowing for more effective treatments of genetic disorders. Unlike Cas9, which targets DNA, Cas13 targets RNA, providing unique advantages for gene modulation. Built on RNA-FM, the modified

DeepFM-Crispr is a novel deep learning model developed to predict the on-target efficiency and assess the off-target effects of Cas13d. This model leverages a large language model to generate detailed representations enriched with evolutionary and structural data, thereby improving predictions of RNA secondary structures and enhancing the overall efficacy of sgRNA [112].

These advancements highlight the critical role of integrating computational predictions into the therapeutic development process, facilitating innovative solutions for both infectious and genetic diseases.

Multi-Species Disease Model Construction

Building multi-species disease models is crucial for understanding the evolutionary mechanisms underlying human diseases. By predicting gene functions across different species, researchers can gain valuable insights into conserved biological mechanisms of disease.

This comparative approach is particularly useful for studying zoonotic diseases and understanding how pathogens evolve, adapt, and spread in various hosts. Utilizing animal and plant models for gene function prediction enables researchers to explore the genetic basis of diseases and develop effective interventions. Machine learning algorithms trained to predict the regulatory activity of nucleic acid sequences have uncovered key principles of gene regulation and guided the analysis of genetic variation. Leveraging machine learning, several novel and powerful approaches have been developed to apply mouse regulatory models for analyzing human genetic variants linked to molecular phenotypes and diseases. These techniques enable the use of thousands of non-human epigenetic and transcriptional profiles, facilitating a more effective exploration of how gene regulation influences human disease [113]. Furthermore, insights from the long-term follow-up of ectopic ACTH-secreting pituitary adenoma emphasize the critical role of extended timeframes in understanding disease progression [114]. Similarly, in the field of multi-species disease modeling, gene prediction models require large-scale datasets and multi-layered training to improve cross-species applicability and stability. Just as case studies and systematic reviews on ectopic thyrotropin-secreting pituitary adenoma provide a foundation for refining clinical insights [115], deep learning-driven disease models benefit from diverse genomic datasets, allowing them to capture species-specific regulatory patterns and functional elements. The integration of deep learning with cross-species genomic data has enabled researchers to construct robust multi-species disease models, improving the identification of conserved pathways and species-specific genetic variations linked to complex diseases. These advances not only deepen our knowledge of pituitary adenomas but also contribute to the development of more accurate cross-species translational research frameworks, fostering advancements in precision medicine and bioinformatics.

Moreover, AI-driven RNA and DNA prediction tools can help identify potential therapeutic targets and aid in the development of cross-species treatments, spanning not only animals but also plants, microorganisms, and humans. This ultimately strengthens our ability to address complex diseases more effectively.

Ethical and Social Issues

The rapid development of gene structure prediction and its application in disease research will bring great convenience and benefits but also raise major ethical and social issues [116]. Potential issues such as data privacy, informed consent, and genetic discrimination must be addressed to ensure responsible development and implementation of these technologies.

In addition, the impact of gene editing technology, especially in terms of its accessibility and potential for abuse, requires a sound ethical framework and regulatory oversight. Ethical considerations often focus on Safety, Informed Consent, Justice and Equity, and Legal aspects (specifically regarding Genome-Editing Research Involving Embryos). The principle of genomic solidarity, with an emphasis on the public good, should serve as a framework for clarifying CRISPR debates. The legitimate claim of genetic exceptionalism highlights the trans-generational risks and helps bridge the knowledge gap [117].

As we address these challenges, promoting public discussion and interdisciplinary collaboration will be critical to shaping the future of gene structure prediction in a balance that prioritizes both ethical considerations and societal benefits.

Future Development Directions

Optimization of Models

The future of gene structure prediction hinges on the optimization of computational models that seamlessly integrate diverse data types, including genomic, transcriptomic, and epigenetic information, enhancing the ability to analyze unannotated data will also be crucial, allowing for more accurate predictions and uncovering new insights into gene function and regulation.

The significance of epigenetics, which refers to inheritance through mechanisms other than the genomic DNA sequence, has been highlighted by its reported involvement in the deregulation of epigenetic processes in human cancer. With the advancement of machine learning and deep learning, which enables multimodal analysis of large omics datasets, it is crucial to develop a platform that can perform multimodal analysis of medical big data using artificial intelligence. This could significantly accelerate the realization of precision medicine [116]. By harnessing advancements in deep learning and artificial intelligence, researchers can create more sophisticated models that better capture the complexity of biological systems, thereby improving the accuracy and reliability of predictions.

Large-Scale Applications

As the field progresses, the focus will shift towards the large-scale application of gene structure prediction in various domains, including drug discovery and personalized genomics. Developing tools that facilitate the identification of disease-associated genes and support personalized genomic research will be paramount [118]. This will involve the creation of personal and user-friendly platforms that allow researchers and clinicians to access and utilize predictive models effectively, this, in turn, accelerates the translation of research findings into clinical practice. Development of tools to support personalized genomics research that will improve healthcare, extend life expectancy, and boost the economy.

Global Collaboration Potential

The potential for global collaboration in gene structure prediction research is immense. Cross-disciplinary partnerships among geneticists, bioinformaticians, and clinicians can drive innovation and enhance our understanding of complex diseases. Collaborative efforts will also facilitate the sharing of data and resources, enabling researchers to tackle pressing health challenges on a global scale.

For example, COVID-19 vaccines were developed at an extraordinary pace, yet international cooperation has seemingly fallen short in ensuring the global equitable distribution of vaccines. A cooperative redistribution scheme, recently launched by WHO, CEPI, and Gavi, aims to address issue [119]. Cross-disciplinary collaboration promotes the advancement of genomic research technology. By fostering a culture of collaboration and knowledge exchange, the scientific community can harness the full potential of gene structure prediction to improve health outcomes worldwide.

Conclusion

The integration of AI models into bioinformatics, especially in gene structure prediction, has driven significant advancements, greatly improved the interpretation of biological data and made a substantial impact on both biomedical research and bioinformatics.

Reflecting on these developments, they have not only enhanced our understanding of genomic architecture but also provided essential insights into the molecular mechanisms behind various diseases. The synergy between computational methods and high-throughput sequencing technologies has enabled gene structure identification and annotation with unprecedented accuracy and efficiency. However, it is crucial to maintain a balanced and objective perspective when interpreting diverse research findings. While advances in computational models, incorporating machine learning and deep learning techniques, have greatly increased predictive capabilities, their biological validity must still be rigorously verified through experimental methods. The mutual reinforcement of computational predictions, empirical evidence, and experimental validation is essential for producing reliable gene annotations that can guide clinical applications and drive new drug development.

Looking to the future, there is a pressing need to further enhance gene structure prediction methods by leveraging smarter and more advanced approaches. This approach combines data from various omics fields, such as genomics, transcriptomics, proteomics, and epigenomics, offering a more comprehensive and detailed understanding of gene function and regulatory mechanisms. By examining these interconnected biological layers, researchers can gain a deeper understanding of how genes are expressed, modified, and regulated in both normal physiological conditions and disease states. Such an integrated strategy holds immense promises for improving predictive accuracy and identifying novel therapeutic targets. Yet, translating these predictions into clinical practice remains a significant challenge. Future efforts must prioritize the development of standardized protocols for validating and applying gene structure predictions in disease diagnosis and treatment, alongside robust legal and ethical frameworks to facilitate their adoption. Collaboration among computational biologists, clinicians, and geneticists will be pivotal in bridging the gap between research insights and their practical application in personalized medicine.

In summary, while advancements in gene structure prediction have opened new avenues for understanding genetic diseases and developing targeted therapies, a balanced, interdisciplinary, and ethically grounded approach is essential to fully unlock the potential of these innovations. By fostering collaboration, ensuring rigorous validation, and adhering to ethical principles, we can continue to improve disease prediction and treatment strategies, ultimately advancing precision medicine, improving patient outcomes, and delivering broader societal benefits.

Discussion

With the booming development of biotechnology, it has brought convenience to people's lives, and at the same time, it has brought technology to the intersection of science and ethics. Technological development is affected by ethical values, and ethical values will affect the development of technology [120]. Although many aspects of biotechnology can be praised for their benefits to mankind. Biotechnology also has its dark side and side effects, which may have unexpected consequences due to improper use, causing harm to humans and society. Therefore, there must be a sound review system to counter the ethical impact of its development.

The broader implications of deep learning and model development in biological research extend beyond technical advancements, touching on significant ethical concerns. One of the critical aspects where deep learning has shown promise is in maintaining privacy and ethical standards in clinical studies. For example, federated learning has emerged as an effective approach for analyzing sensitive biological data while maintaining privacy and security. Zhang et al. demonstrated the effectiveness of federated learning in predicting postoperative remission in patients with acromegaly across multiple medical centers, all while preserving patient confidentiality [121]. This study highlights how distributed machine learning approaches can mitigate the risks associated with centralized data storage and potential breaches, thereby promoting ethical standards in biomedical research.

Every major advancement in technology has the potential to profoundly impact the world. As such, the ethical evaluation of emerging technologies, including biotechnology, must evolve in step with these advancements. Each technological breakthrough not only brings convenience to public life but also introduces unique ethical challenges that demand flexible responses from policymakers and legal experts based on real-world conditions. Given the profound implications for the future trajectory of human development, these issues must be approached with careful deliberation. Beyond safeguarding rights and freedoms, we must exercise wisdom to anticipate and mitigate potential risks. Genetic privacy, for example, is a major issue as genomic data contains highly sensitive information that can reveal personal health risks and ancestral background. Ensuring the anonymity and security of such data is crucial. Deep learning models must integrate privacy-preserving techniques, such as differential privacy and homomorphic encryption, to effectively address these challenges. Another crucial ethical consideration is equitable access to advanced technologies. As deep learning models become increasingly sophisticated, there is a risk of widening the gap between well-funded institutions and under-resourced regions. Ensuring that these cutting-edge technologies are accessible to researchers and clinicians worldwide is essential to avoid exacerbating existing disparities in healthcare and research capabilities. Policies and initiatives that promote open-source model sharing and capacity-building programs can help bridge this gap and democratize access to deep learning advancements in biology.

Furthermore, addressing biases in training datasets is critical to ensure fair and unbiased model performance across diverse populations. Many deep learning models are trained on datasets that may not fully represent the genetic diversity of global populations, leading to biased predictions and potential misdiagnoses. Ethical frameworks should emphasize the importance of inclusive data collection strategies and transparency in model development to mitigate these biases and enhance trust in AI-driven biomedical applications. Compared to other fields, deep learning offers unprecedented opportunities in biological research. Addressing ethical challenges such as genetic privacy, equitable access, and dataset biases is crucial to ensure responsible and fair implementation. Future research should prioritize the development of ethical guidelines and regulatory frameworks to govern the use of deep learning in genomics and healthcare, ensuring a balance between innovation and ethical responsibility.

Additionally, the development of biotechnology is inextricably linked to finance. Biotechnology innovation relies on significant capital investment, and its growth can, in turn, generate employment opportunities and drive economic progress. The commercialization of biotechnological advancements, such as novel drug development, precision medicine, and agricultural biotechnology, requires substantial funding for research, regulatory approvals, and market deployment. Public and private investment in biotechnology not only fosters scientific breakthroughs but also contributes to the economic ecosystem by creating high-skilled jobs, attracting foreign investment, and boosting industrial competitiveness.

Governments and financial institutions play a crucial role in supporting biotech startups and research institutions through funding initiatives, tax incentives, and favorable regulatory environments. As the biotech sector continues to expand, collaboration between academia, industry, and financial stakeholders will be essential to sustain growth and ensure that the benefits of biotechnological innovation are widely accessible.

In summary, the advancement of biotechnology creates a ripple effect, extending far beyond mere technological progress to significantly influence human society. This underscores the importance of establishing sound laws and policies to ensure that its development fosters a positive and sustainable cycle of growth. Biotechnology not only transforms healthcare and agriculture but also shapes economic structures, educational frameworks, and ethical norms. Effective regulatory frameworks and strategic investments are essential to maximize its benefits while mitigating potential risks. In the evolution of human society, no discipline operates in isolation; progress is built on the interconnected contributions of multiple fields. Collaboration between scientific research, financial investments, policy-making, and public engagement is crucial to harnessing the full potential of biotechnological advancements. By fostering a holistic and inclusive approach, society can ensure that biotechnology continues to drive innovation while promoting ethical, social, and economic well-being on a global scale.

Author Contributions

Tong Wang did the supervision, visualization, and writing of the manuscript. Jing-Min Yang did the conceptualization, writing and outline the manuscript. Ming Wu did the supervision and organization of the manuscript. Ting Xu did data curation, writing and bioinformation support. Yuanyin Teng and Yuqing Miao did the figure detailed drawing support.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Reference

1. Bayat A. Science, medicine, and the future: Bioinformatics. *British Medical Journal.* 2002;324(April):1018-1022. Available at: www.ebi.ac.uk

2. Miao T, Guo J. Application of artificial intelligence deep learning in numerical simulation of seawater intrusion. *Environmental Science and Pollution Research.* 2021;28(38):54096-54104. doi:10.1007/s11356-021-13680-5

3. Liu J, Li J, Wang H, Yan J. Application of deep learning in genomics. *Science China Life Sciences.* 2020;63(12):1860-1878. doi:10.1007/s11427-020-1804-5

4. Kim P, Tan H, Liu J, Yang M, Zhou X. FusionAI: Predicting fusion breakpoint from DNA sequence with deep learning. *iScience.* 2021;24(10):103164. doi:10.1016/j.isci.2021.103164

5. Jagodnik KM, Shvili Y, Bartal A. HetIG-PreDiG: A Heterogeneous Integrated Graph Model for Predicting Human Disease Genes based on gene expression. *PLoS One.* 2023;18(2):1-27. doi:10.1371/journal.pone.0280839

6. Yousef M, Allmer J. Deep learning in bioinformatics. *Turkish Journal of Biology.* 2023;47(6):366-382. doi:10.55730/1300-0152.2671

7. Ahmad RM, Ali BR, Al-Jasmi F, Sinnott RO, Al Dhaheri N, Mohamad MS. A review of genetic variant databases and machine learning tools for predicting the pathogenicity of breast cancer. *Briefings in Bioinformatics.* 2024;25(1):1-20. doi:10.1093/bib/bbad479

8. Liu X, Zhang H, Zeng Y, Zhu X, Zhu L, Fu J. DRANetSplicer: A Splice Site Prediction Model Based on Deep Residual Attention Networks. *Genes.* 2024;15(4):40404. doi:10.3390/genes15040404

9. Parvez MR, Hu W, Chen T. Comparison of the Smith-Waterman and Needleman-Wunsch algorithms for online similarity analysis of industrial alarm floods. *2020 IEEE Electric Power and Energy Conference (EPEC).* 2020;0(0):1-6. doi:10.1109/EPEC48502.2020.9320080

10. Ma J, Qin T, Xiang J. Disease-gene prediction based on preserving structure network embedding. *Frontiers in Aging Neuroscience.* 2023;15:1061892. doi:10.3389/fnagi.2023.1061892

11. Drukewitz SH, von Reumont BM. The significance of comparative genomics in modern evolutionary venomics. *Frontiers in Ecology and Evolution.* 2019;7(May):163. doi:10.3389/fevo.2019.00163

12. Dalal A, Atri A. An introduction to sequence and series. *International Journal of Research.* 2014;1(10):1286-1292. doi:10.1002/0471250953.bi0301s42.an

13. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution.* 1987;25(4):351-360. doi:10.1007/BF02603120

14. Nagaraj P. New sequence alignment algorithm using AI rules and dynamic seeds. *Bioscience Engineering: An International Journal.* 2023;10(1-2):1-14. doi:10.5121/bioej.2023.10201

15. Reddy B, Fields R. Multiple sequence alignment algorithms in bioinformatics. *Lecture Notes in Networks and Systems.* 2022;286(February):89-98. doi:10.1007/978-981-16-4016-2_9

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology.* 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

17. Mohamed EM, Mousa HM, Keshk AE. Comparative analysis of multiple sequence alignment tools. *International Journal of Information Technology and Computer Science.* 2018;10(8):24-30. doi:10.5815/ijitcs.2018.08.04

18. Ocaña K, Oliveira F, Dias J, Ogasawara E, Mattoso M. Designing a parallel cloud-based comparative genomics workflow to improve phylogenetic analyses. *Future Generation Computer Systems.* 2013;29:2205-2219. doi:10.1016/j.future.2013.04.005

19. Taniguchi T, Okuno M, Shinoda T, et al. GINGER: An integrated method for high-accuracy prediction of gene structure in higher eukaryotes at the gene and exon level. *DNA Research.* 2023;30(4):dsad017. doi:10.1093/dnares/dsad017

20. Rosati D, Palmieri M, Brunelli G, et al. Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal.* 2024;23(February):1154-1168. doi:10.1016/j.csbj.2024.02.018

21. Hassan M, Awan FM, Naz A, et al. Innovations in genomics and big data analytics for personalized medicine and health care: A review. *International Journal of Molecular Sciences.* 2022;23(9):94645. doi:10.3390/ijms23094645

22. Cen LP, Ng TK, Ji J, et al. Artificial intelligence-based database for prediction of protein structure and their alterations in ocular diseases. *Database.* 2023;2023:1-10. doi:10.1093/database/baad083

23. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74. doi:10.1038/nature11247

24. Chen Z, Ain NU, Zhao Q, Zhang X. From tradition to innovation: conventional and deep learning frameworks in genome annotation. *Briefings in Bioinformatics.* 2024;25(3):bbad138. doi:10.1093/bib/bbae138

25. Hoheisel JD. Application of hybridization techniques to genome mapping and sequencing. *Trends in Genetics*. 1994;10(3):79-83. doi:10.1016/0168-9525(94)90229-1

26. Xiang J, Meng X, Zhao Y, Wu FX, Li M. HyMM: hybrid method for disease-gene prediction by integrating multiscale module structure. *Briefings in Bioinformatics.* 2022;23(3):bbac072. doi:10.1093/bib/bbac072

27. Hüttenhofer A, Vogel J. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research.* 2006;34(2):635-646. doi:10.1093/nar/gkj469

28. Harrow J, Nagy A, Reymond A, et al. Identifying protein-coding genes in genomic sequences. *Genome Biology.* 2009;10(1):201. doi:10.1186/gb-2009-10-1-201

29. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;21(18):3674-3676. doi:10.1093/bioinformatics/bti610

30. Zdobnov EM, Apweiler R. InterProScan - An integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 2001;17(9):847-848. doi:10.1093/bioinformatics/17.9.847

31. Besemer J, Borodovsky M. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research.* 2005;33(SUPPL. 2):451-454. doi:10.1093/nar/gki487

32. Dubchak I, Frazer K. Multi-species sequence comparison: The next frontier in genome annotation. *Genome Biology.* 2003;4(12):122. doi:10.1186/gb-2003-4-12-122

33. Berger B, Yu YW. Navigating bottlenecks and trade-offs in genomic data analysis. *Nature Reviews Genetics.* 2023;24(4):235-250. doi:10.1038/s41576-022-00551-z

34. Sharma S, Chaudhary P. Machine learning and deep learning. *Quantum Computing Artificial*

Intelligence: Training Machine Deep Learning Algorithms. Published online 2023:71-84. doi:10.1515/9783110791402-004

35. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444. doi:10.1038/nature14539

36. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. *Human Genomics.* 2022;16(1):26. doi:10.1186/s40246-022-00396-x

37. Koumakis L. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*. 2020;18:1466-1473. doi:10.1016/j.csbj.2020.06.017

38. Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics.* 2021;37(15):2112-2120. doi:10.1093/bioinformatics/btab083

39. Hallee L, Rafailidis N, Gleghorn JP. cdsBERT - Extending protein language models with codon awareness. *bioRxiv.* Published online 2023. doi:10.1101/2023.09.15.558027

40. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics.* 2022;38(8):2102-2110. doi:10.1093/bioinformatics/btac020

41. Zhou Z, Ji Y, Li W, Dutta P, Davuluri RV, Liu H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. *The Twelfth International Conference on Learning Representations.* 2024. Available at: https://openreview.net/forum?id=oMLQB4EZE1

42. Feng H, Wu L, Zhao B, et al. Benchmarking DNA foundation models for genomic sequence classification. *bioRxiv Preprint Server for Biology.* Published online 2024. doi:10.1101/2024.08.16.608288

43. Koroteev MV. BERT: A review of applications in natural language processing and understanding. *arXiv*. 2021 ;(March). doi:10.48550/arXiv.2103.11943

44. Mohiuddin K, Welke P, Alam MA, et al. Retention is all you need. International Conference on Information and Knowledge Management Proceedings. 2023;(Nips):4752-4758. doi:10.1145/3583780.3615497

45. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies Proceedings. 2019; 1(Mlm):4171-4186.

46. Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. Large language models in bioinformatics: applications and perspectives. *ArXiv*. Published online 2024.
Available at: http://www.ncbi.nlm.nih.gov/pubmed/38259343

47. Sarumi OA, Heider D. Large language models and their applications in bioinformatics. *Computational and Structural Biotechnology Journal.* 2024;23(September):3498-3505. doi:10.1016/j.csbj.2024.09.031

48. Zhou Z, Wu W, Ho H, et al. DNABERT-S: Learning species-aware DNA embedding with genome foundation models. *arXiv.* Published online 2024:1-21.
Available at: http://arxiv.org/abs/2402.08777

49. Perez R, Li X, Giannakoulias S, Petersson EJ. AggBERT: Best in class prediction of hexapeptide amyloidogenesis with a semi-supervised ProtBERT model. *Journal of Chemical Information and Modeling.* 2023;63(18):5727-5733. doi:10.1021/acs.jcim.3c00817

50. Li Y, Ngom A. Data integration in machine learning. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* 2015:1665-1671. doi:10.1109/BIBM.2015.7359925

51. Noble WS. Support vector machine applications in computational biology. *Methods in Molecular Biology.* Published online 2004;71-92.

52. Chiang M, Brackley CA, Naughton C, et al. Genome-wide chromosome architecture prediction reveals biophysical principles underlying gene structure. *Cell Genomics.* 2024;4(12):100698. doi:10.1016/j.xgen.2024.100698

53. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature.* 2014;505(7485):696-700. doi:10.1038/nature12756

54. Yang X, Yang M, Deng H, Ding Y. New era of studying RNA secondary structure and its influence on gene regulation in plants. *Frontiers in Plant Science.* 2018;9(May):671. doi:10.3389/fpls.2018.00671

55. Chen J, Hu Z, Sun S, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *bioRxiv.* Published online 2022:1-23. doi:10.1101/2022.04.03.00300

56. Yu H, Yang H, Sun W, et al. PlantRNA-FM: An interpretable RNA foundation model for exploration functional RNA motifs in plants. *bioRxiv.* Published online 2024:2024.06.24.600509. doi:10.1038/s42256-024-00946-z

57. Yu H, Yang H, Sun W, et al. An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nature Machine Intelligence.* 2024;6(12):1616-1625. doi:10.1038/s42256-024-00946-z

58. Zhang X, Zhou X, Wan M, Xuan J, Jin X, Li S. PINC: A tool for non-coding RNA identification in plants based on an automated machine learning framework. *International Journal of Molecular Sciences.* 2022;23(19):11825. doi:10.3390/ijms231911825

59. Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017; 2017-Decem(Nips):5999-6009. doi:10.5555/3295222.3295349

60. Rader B, Hswen Y, Brownstein JS. Further reflections on the use of large language models in pediatrics. *JAMA Pediatrics.* 2024;178(6):628-629. doi:10.1001/jamapediatrics.2024.0729

61. Xing J, Xing R, Sun Y. Comparative analysis of pooling mechanisms in LLMs: A sentiment analysis perspective. *arXiv.* Published online 2024. Available at: https://arxiv.org/abs/2411.14654

62. Xing J, Xing R, Sun Y. FGATT: A robust framework for wireless data imputation using fuzzy graph attention networks and transformer encoders. *arXiv.* Published online 2024. Available at: https://arxiv.org/abs/2412.01979

63. Gorenstein L, Konen E, Green M, Klang E. BERT in radiology: A systematic review of natural language processing applications. *Journal of the American College of Radiology.* 2024;21(6):914-941. doi:10.1016/j.jacr.2024.01.012.

64. Noh J, Kavuluru R. Improved biomedical word embeddings in the transformer era. *Journal of Biomedical Informatics.* 2021;120:103867. doi:10.1016/j.jbi.2021.103867

65. Song T, Song H, Pan Z, Gao Y, Dai H, Wang X. DeepDualEnhancer: A dual-feature input DNABert-based deep learning method for enhancer recognition. *International Journal of Molecular Sciences.* 2024;25(21):11744. doi:10.3390/ijms252111744

66. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: Considerations in platform selection, data preprocessing and analysis. *Epigenomics.*

2015;7(5):813-828. doi:10.2217/epi.15.21

67. Van R, Alvarez D, Mize T, et al. A comparison of RNA-Seq data preprocessing pipelines for transcriptomic predictions across independent studies. *BMC Bioinformatics.* 2024;25(1):1-22. doi:10.1186/s12859-024-05801-x

68. Wang G, Li S, Yan Q, et al. Optimization and evaluation of viral metagenomic amplification and sequencing procedures toward a genome-level resolution of the human fecal DNA virome. *Journal of Advanced Research.* 2023;48:75-86. doi:10.1016/j.jare.2022.08.011

69. Fukasawa Y, Ermini L, Wang H, Carty K, Cheung MS. LongQC: A quality control tool for third generation sequencing long read data. *G3: Genes, Genomes, Genetics.* 2020;10(4):1193-1196. doi:10.1534/g3.119.400864

70. Deciphering the impact of genomic variation on function. *Nature*. 2024;633(8028):47-57. doi:10.1038/s41586-024-07510-0

71. Slavotinek A, Prasad H, Yip T, Rego S, Hoban H, Kvale M. Predicting genes from phenotypes using human phenotype ontology (HPO) terms. *Human Genetics.* 2022;141(11):1749-1760. doi:10.1007/s00439-022-02449-6

72. Zhao M, Havrilla JM, Fang L, et al. Phen2Gene: Rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics and Bioinformatics.* 2020;2(2):1-12. doi:10.1093/nargab/lqaa032

73. Gokhman D, Harris KD, Carmi S, Greenbaum G. Predicting the direction of phenotypic difference. *bioRxiv*. Published online 2024:2024.02.22.581566. doi:10.1101/2024.02.22.581566

74. Zhou J, Wong MS, Chen WC, Krainer AR, Kinney JB, McCandlish DM. Higher-order epistasis and phenotypic prediction. *Proceedings of the National Academy of Sciences of the United States of America.* 2022;119(39):e2204233119. doi:10.1073/pnas.2204233119

75. Campagne F, Dorff KC, Chambwe N, Robinson JT, Mesirov JP. Compression of structured high-throughput sequencing data. *PLoS One*. 2013;8(11). doi:10.1371/journal.pone.0079871

76. Marchet C, Boucher C, Puglisi SJ, Medvedev P, Salson M, Chikhi R. Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Research.* 2021;31(1):1-12. doi:10.1101/gr.260604.119

77. Lian H, Xiong Y, Niu J, et al. Scaffold-BPE: Enhancing Byte Pair Encoding with Simple and

Effective Scaffold Token Removal. *bioRxiv*. Published online 2022. Available at: https://arxiv.org/abs/2404.17808

78. Kumar S, Singh MP, Nayak SR, et al. A new efficient referential genome compression technique for FastQ files. *Functional and Integrative Genomics*. 2023;23(4):333. doi:10.1007/s10142-023-01259-x

79. Md Riyad Hossain, Dr. Douglas Timmer. Machine Learning Model Optimization with Hyper Parameter Tuning Approach. *Global Journal of Computer Science and Technology*. 2021;21(2):1-1. Available at: https://core.ac.uk/download/pdf/539593628.pdf

80. Frazier PI. A Tutorial on Bayesian Optimization. 2018;(Section 5):1-22. Available at: http://arxiv.org/abs/1807.02811 doi:10.48550/arXiv.1807.02811

81. Li ZY, Gao S, Cheng MM. SERE: Exploring Feature Self-Relation for Self-Supervised Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(12):15619-15631. doi:10.1109/TPAMI.2023.3309979

82. Vojgani E, Pook T, Martini JWR, et al. Accounting for epistasis improves genomic prediction of phenotypes with univariate and bivariate models across environments. *Theoretical and Applied Genetics*. 2021;134(9):2913-2930. doi:10.1007/s00122-021-03868-1

83. Vojgani E, Pook T, Simianer H. Phenotype Prediction Under Epistasis BT - Epistasis: Methods and Protocols. In: Wong K-C, ed. Springer US; 2021:105-120. doi:10.1007/978-1-0716-0947-7_8

84. Richter T, Bahrami M, Xia Y, Fischer DS, Theis FJ. Delineating the Effective Use of Self-Supervised Learning in Single-Cell Genomics. *bioRxiv*. Published online 2024:2022-2024. doi:10.1038/s42256-024-00934-3

85. Gündüz HA, Binder M, To XY, et al. A self-supervised deep learning method for data-efficient training in genomics. *Communications Biology*. 2023;6(1):1-12. doi:10.1038/s42003-023-05310-2

86. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*. 2019;20(7):389-403. doi:10.1038/s41576-019-0122-6

87. Press O, Smith NA, Lewis M. Train Short, Test Long: Attention With Linear Biases Enables Input Length Extrapolation. *ICLR 2022 - 10th International Conference on Learning*

Representations. Published online 2022. Available at: https://arxiv.org/abs/2202.11256

88. Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*. 2002;35(2):142-150. doi:10.1016/S1532-0464(02)00506-3

89. Gupta P, Vishnudas CK, Robin V V., Dharmarajan G. Host phylogeny matters: Examining sources of variation in infection risk by blood parasites across a tropical montane bird community in India. *Parasites and Vectors*. 2020;13(1):1-13. doi:10.1186/s13071-020-04404-8

90. Caraza-Harter M V, Endelman JB. The genetic architectures of vine and skin maturity in tetraploid potato. *Theoretical and Applied Genetics*. 2022;135(9):2943-2951. doi:10.1007/s00122-022-04159-z

91. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv Prepr arXiv230615006*. Published online 2023. doi:10.48550/arXiv.2306.15006

92. Wang K, Zeng X, Zhou J, Liu F, Luan X, Wang X. BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning. *Briefings in Bioinformatics.* 2024;25(3):bbae195. doi:10.1093/bib/bbae195

93. Zhang J, Fei Y, Sun L, Zhang QC. Advances and opportunities in RNA structure experimental determination and computational modeling. *Nat Methods*. 2022;19(10):1193-1207. doi:10.1038/s41592-022-01623-y

94. Ren Z, Jiang L, Di Y, et al. CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics*. 2024;40(7):btae330. doi:10.1093/bioinformatics/btae330

95. Xiang Y, Huang W, Tan L, et al. Pervasive downstream RNA hairpins dynamically dictate start-codon selection. *Nature*. 2023;621(7978):423-430. doi:10.1038/s41586-023-06500-y

96. Kabir A, Bhattarai M, Peterson S, et al. DNA breathing integration with deep learning foundational model advances genome-wide binding prediction of human transcription factors. *Nucleic Acids Research*. 2024;52(19):e91. doi:10.1093/nar/gkae783

97. Kabir A, Bhattarai M, Peterson S, Najman-Licht Y, Rasmussen KØ, Shehu A, Bishop AR, Alexandrov B, Usheva A. DNA breathing integration with deep learning foundational model advances genome-wide binding prediction of human transcription factors. Nucleic Acids

Research. 2024;52(19):e91. doi:10.1093/nar/gkae783

98. Wei J, Chen S, Zong L, Gao X, Li Y. Protein–RNA interaction prediction with deep learning : *Briefings in Bioinformatics.* Published online 2022:1-19. doi:10.1093/bib/bbab098

99. Zhu J, Wang Z, Zhang Y, Li X, Liu J, Deng K, Lu L, Pan H, Wang R, Yao Y. Ectopic pituitary adenomas: clinical features, diagnostic challenges and management. *Pituitary.* 2023;23(6):648-664. doi:10.1007/s11102-023-01234-5

100. Li X, Deng K, Zhang Y, Feng M, Xing B, Lian W, Yao Y. Pediatric pituitary neuroendocrine tumors—a 13-year experience in a tertiary center. *Frontiers in Oncology.* 2023;13(4):1270958. doi:10.3389/fonc.2023.1270958

101. Wellenreuther M, Mérot C, Berdan E, Bernatchez L. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology.* 2019;28(6):1203-1209. doi:10.1111/mec.15066

102. Kabir A, Bhattarai M, Peterson S, Najman-Licht Y, Rasmussen KØ, Shehu A, Bishop AR, Alexandrov B, Usheva A. DNA breathing integration with deep learning foundational model advances genome-wide binding prediction of human transcription factors. Nucleic Acids Research. 2024;52(19):e91. doi:10.1093/nar/gkae783

103. Avramouli A, Krokidis MG, Exarchos TP, Vlamos P. Protein structure prediction for disease-related insertions/deletions in Presenilin 1 gene. GeNeDis 2022. Springer International Publishing; 2023:31-40. doi:10.1007/978-3-031-09547-0_3

104. Wu J, Ouyang J, Qin H, et al. PLM-ARG: Antibiotic resistance gene identification using a pretrained protein language model. Bioinformatics. 2023;39(11):btad690. doi:10.1093/bioinformatics/btad690

105. Caprara MG, Nilsen TW. RNA: Versatility in form and function. Nature Structural Biology. 2000;7(10):831-833. doi:10.1038/82816

106. Laberge A, Burke W. Personalized Medicine and Genomics. *Hastings Center Bioethics Briefings.* 2008;(May):133-136. Available at: https://bioethics.org.gr/wp-content/uploads/2024/04/personalizedmedicinehastings.pdf

107. Wu Q, Boueiz A, Bozkurt A, et al. Deep Learning Methods for Predicting Disease Status Using Genomic Data HHS Public Access. *J Biom Biostat.* 2018;9(5). *Journal of Biometrics and Biostatistics.* 2018;9(5):417. doi:10.4172/2155-6180.1000417

108. Mohammed Aarif KO, Mohammed Yousuf Hasan V, Alam A, Shoukath Ali K, Pakruddin B. Chapter 4 - Decoding DNA: Deep learning's impact on genomic exploration. In: Raza KBT-DL in G and G, ed. Academic Press; 2025:77-95. doi:https://doi.org/10.1016/B978-0-443-27574-6.00005-9

109. Chuai G, Ma H, Yan J, et al. DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. *Genome Biology.* 2018;19(1):1-19. doi:10.1186/s13059-018-1459-4

110. Zhang S, Zhou J, Hu H, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Research.* 2015;44(4):1-14. doi:10.1093/nar/gkv1025

111. Li Y, Umbach DM, Krahn JM, Shats I, Li X, Li L. Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC Genomics*. 2021;22(1):272. doi:10.1186/s12864-021-07581-7

112 Bao C, Liu F. DeepFM-Crispr: Prediction of CRISPR On-Target Effects via Deep Learning. *arXiv.* Published online 2024. Available at: http://arxiv.org/abs/2409.05938

113. Kelley DR. Cross-species regulatory sequence activity prediction. *PLoS Computational Biology.* 2020;16(7):1-27. doi:10.1371/journal.pcbi.1008050

114. Zhu J, Lu L, Yao Y, Chen S, Li W, You H, Feng F, Feng M, Zhang Y.Long-term follow-up for ectopic ACTH-secreting pituitary adenoma in a single tertiary medical center and a literature review. *Pituitary.* 2023;23(2):149-159. doi:10.1007/s11102-023-01212-w

115. Li X, Zhao B, Hou B, Wang J, Zhu J, Yao Y, Lian X. Case report and literature review: ectopic thyrotropin-secreting pituitary adenoma in the suprasellar region. *Frontiers in Endocrinology.* 2023;12(3):619161. doi:10.3389/fendo.2023.619161

116. Salerno J, Coughlin SS, Goodman KW, Hlaing WM. Current ethical and social issues in epidemiology. *Annals of Epidemiology.* 2023;80:37-42. doi:https://doi.org/10.1016/j.annepidem.2023.02.001

117. Mulvihill JJ, Capps B, Joly Y, Lysaght T, Zwart HAE, Chadwick R. Ethical issues of CRISPR technology and gene editing through the lens of solidarity. *British Medical Bulletin.* 2017;122(1):17-29. doi:10.1093/bmb/ldx002

118. Hamamoto R, Komatsu M, Takasawa K, Asada K, Kaneko S. Epigenetics Analysis and Integrated Analysis of Multiomics Data, Including Epigenetic Data, Using Artificial

Intelligence in the Era of Precision Medicine. *Biomolecules*. 2020;10(1). doi:10.3390/biom10010062

119. Luna F, Holzer F. International cooperation in a non-ideal world: the example of COVAX. *Cadernos Ibero-Americanos de Direito Sanitário*. 2021;10(3):199-210. doi:10.17566/ciads.v10i3.789

120. O'Mathúna DP. Bioethics and biotechnology. *Cytotechnology*. 2007;53(1-3):113-119. doi:10.1007/s10616-007-9053-8

121. Zhang W, Wu X, Wang H, et al. Federated Learning for Predicting Postoperative Remission of Patients with Acromegaly: A Multicentered Study. *World Neurosurg*. 2024;193:1036-1046. doi:10.1016/j.wneu.2024.10.091

FIGURES AND TABLES

Preview: History of Gene Structure Prediction

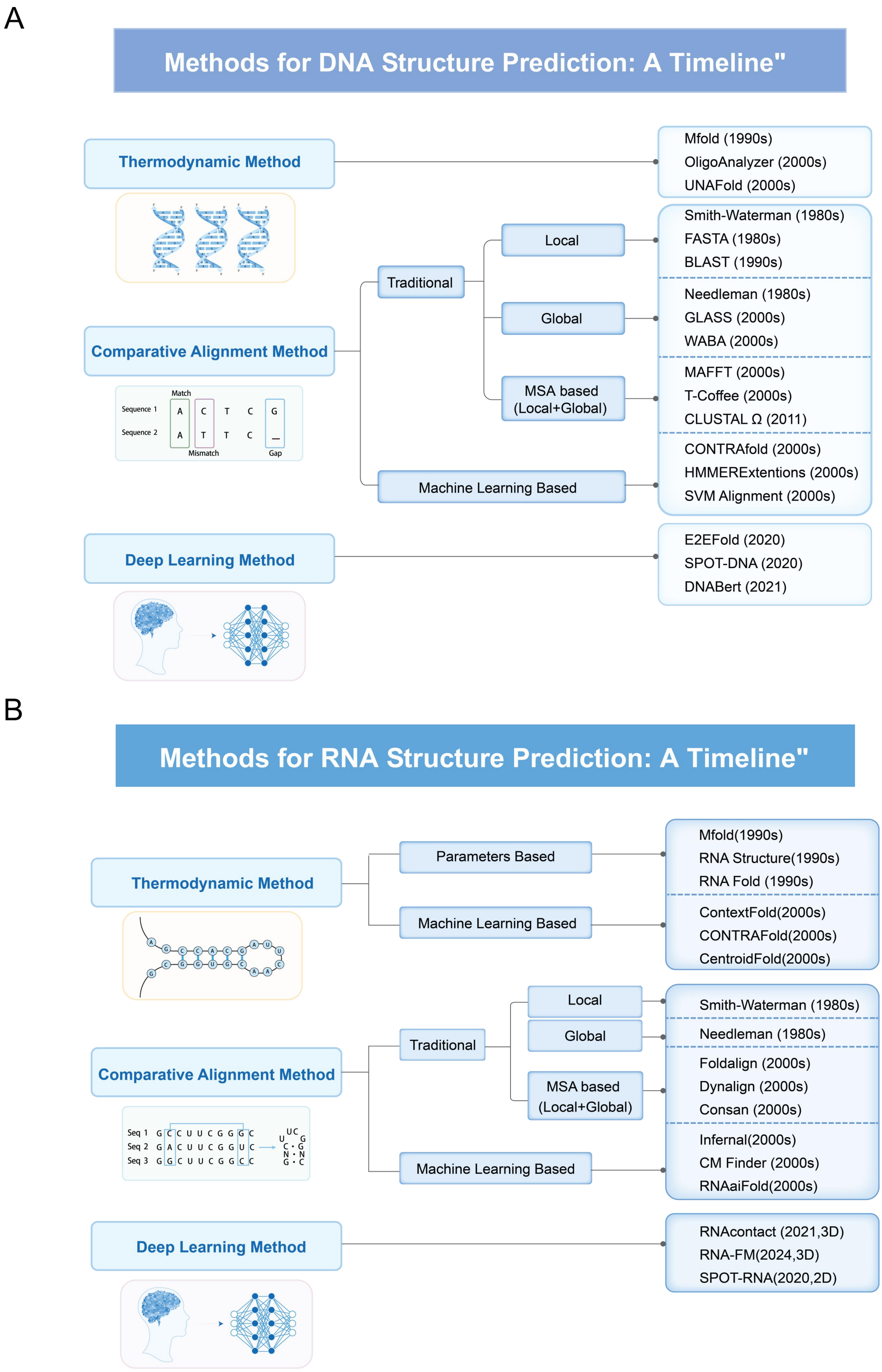

Figure 1. Timeline Advancements in Gene Structure Prediction

(A) Methods for DNA Structure Prediction. This diagram illustrates the evolution of DNA structure prediction methods over time, categorized into three primary approaches: thermodynamic, comparative alignment, and deep learning methods. The thermodynamic method includes local and global approaches, with key milestones such as Mfold (1990s), OligoAnalyzer (2000s), and UNAFold (2000s). Comparative alignment methods have progressed from traditional algorithms, such as Smith-Waterman (1980s), Needleman (1980s) and BLAST (1990s); to multiple sequence alignment (MSA)-based approaches like MAFFT (2000s), T-Coffee (2000s) and Clustal Ω (2011). More recent advancements incorporate machine learning-based approaches including CONTRAfold (2000s), and deep learning models such as E2E-Fold (2020), SPOT-DNA (2020) and DNABert (2021).

(B) Methods for RNA Structure Prediction. This diagram illustrates the evolution of DNA structure prediction methods over time, categorized into three primary approaches: thermodynamic, comparative alignment, and deep learning methods. The thermodynamic approach includes parameter-based methods such as Mfold (1990s), RNA Structure (1990s), and RNA FOLD (1990s), while machine learning-based approaches have gained prominence in recent years including ContextFold (2000s). Comparative alignment methods have evolved from traditional approaches, such as Smith-Waterman (1980s) and Needleman (1980s), to MSA-based approaches including Foldalign (2000s), and to machine learning-based approaches including Inferal (2000s). Recent advances in deep learning-based techniques include RNAContext (2021), RNA-FM (2023), and SPOT-RNA (2020).

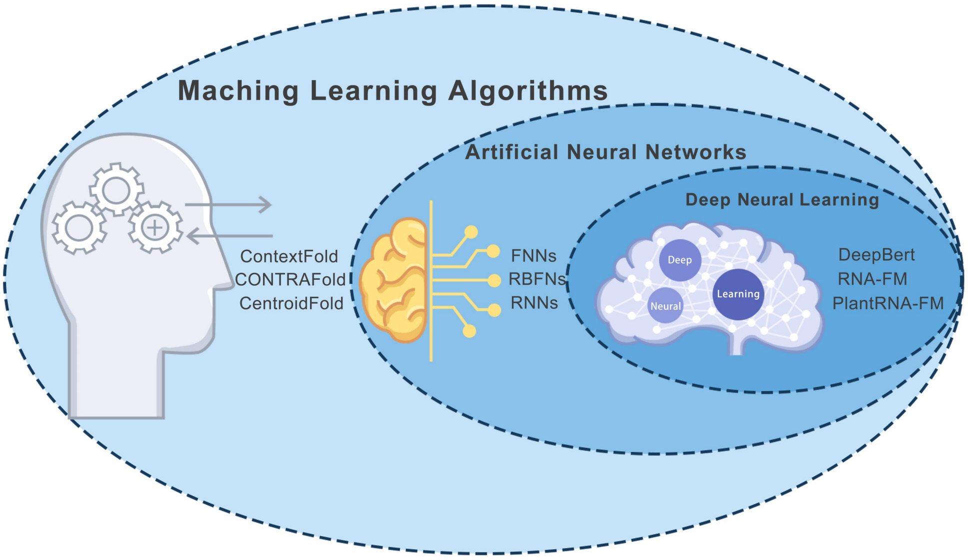

Figure 2. Diagram of AI Machine Learning and Deep Learning Concepts and Classes

This Venn diagram depicts the hierarchical relationship between machine learning algorithms, artificial neural networks (ANNs), and deep learning techniques, particularly in the realm of RNA and DNA structure prediction. The outermost layer represents machine learning algorithms, including tools such as CentroidFold, CONTRAfold, and ContextFold. Within this layer, artificial neural networks (ANNs) are highlighted as a subset, encompassing methodologies like Feedforward Neural Network (FNN), Radial Basis Function Networks (RBFNs), and Recurrent Neural Networks (RNNs). At the core of the diagram is deep neural learning, a specialized branch of ANNs, which includes advanced models such as DeepBert, RNA FM, and ProteinRNA-FM. This figure emphasizes the nested structure of these computational approaches, with deep neural learning offering the most sophisticated modeling capabilities among the three.

TABLE 1. Comparison of MSA techniques

Technique	Input Format	Output Format	Seq Type	Method	Server
CLUSTAL OMEGA	FASTA, EMB, GenBank	ClustalW/ Pearson/FASTA/ MSF	Protein, DNA, RNA	Global/ Progressive	http://www.clustal.org/o mega/ https://www.ebi.ac.uk/T ools/msa/clustalo/
MUSCL	FASTA, EMB, GenBank	Fasta, Clustalw, MSF/html	Protein	Progressive Step1 and Step2 iterative Step 3	http://www.drive5.com/ muscle/ https://www.ebi.ac.uk/Tools/msa/muscle/
MAFFT	FASTA, EMB, GenBank	ClustalW/ Pearson/FASTA	Protein, DNA, RNA	Global/ Iterative	http://mafft.cbrc.jp/alignment/server/ https://www.ebi.ac.uk/Tools/msa/mafft/
KALIGN	FASTA, EMB, GenBank	MACSIM/ ClustalW/ Pearson/FASTA	Protein, DNA, RNA	Progressive	http://msa.sbc.su.se/cgi bin/msa.cgi https://www.ebi.ac.uk/Tools/msa/kalign/
RETALIGN	FASTA	ClustalW	Protein	Progressive Corner cutting Multiple Sequence Alignment	http://phylogenycafe.elte .hu/RetAlign/
PROBCONS	MFA	MFA/ClustalW	Protein	Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences	http://probcons.stanford.Edu

TABLE 2. Deep learning-based genomic tools and algorithms for variant calling and annotation

Tools	DL model	Application	Website Code Source
Clairvoyante	CNN	To predict variant type, zygosity, alternative allele and Indel length	https://github.com/aquaskyline/Clair voyante
DeepVariant	CNN	To call genetic variants from next generation DNA sequencing data	https://github.com/google/deepv ariant
GARFIELD-NGS	DNN + MLP	To classify true and false variants from WES data	https://github.com/gedoardo83/ GARFIELD-NGS
Intelli-NGS	ANN	To define good and bad variants calls from Ion Torrent sequencer data	https://github.com/aditya-88/intel li-ngs
DAVI (Deep Alignment and Variant Identifica tion)	CNN + RNN	To identify variants in NGS reads	N/A
DeepSV	CNN	To call genomic deletions by visualizing sequence reads	https://github.com/CSuperlei/DeepSV

TABLE 3. From left to right the columns represent the DL model

Name	DL model	Omics data	Purpose/Prediction	Accuracy
DeepTarget	RNN	miRNA-mRNA pairing	Target Prediction	0,96
DeepMirGene	LSTM	Positive pre-miRNA and non-miRNA	miRNA Target	0.89 Sensitivity
DeepNet	ANN	RNA-Seq	Control-cases	-0.7
	AE	Time-series Gene Expression	Pre-processing step for Clustering	
	AE	cDNA Microarrays	Predict the Organization of Transcriptomic Machinery	
ADAGE	AE	Gene Expression	Identification/ Reconstruction of Biological Signals	
eADAGE	AE	Gene Expression	Identification of Biological Patterns	
D-GEX	RNN	Expression of Landmark Genes	Gene Expression Inference	Overall, Error 0.3204±0.0879
DeepChrome	CNN	Histone Modifications	Classify Gene Expression	Average area under the curve (AUC)=0.80
AttentiveChrome	LSTM	Histone Modifications	Classify Gene Expression	Average AUC=0.81
Multimodaldeep belief network	DBN	Gene expression, DNA Methylation and miRNA Expression	Identification of Key Genes and miRNAs	Average Correlations 0.91, 0.73 and 0.69 for the GE, DM and ME
DeepVariant	CNN	Whole-genome Sequence	Variant Caller	99,45% F1
	ANN	Cell-line with Drug Response	Predict Drug Response	0.65 AUC
DeepFIGV	CNN	Whole-genome Sequence	Predict Quantitative Epigenetic Variation	z-scores DNase rho=0.0802, P=5.32e 16
DeePathology	Multiple AEs	mRNA and miRNA	Predict Tissue-of-origin, Normal or Disease State and Cancer Type	99.4% Accuracy for Cancer Subtype
DeepCpG	CNN	Single Cell Methylation	Predicts Missing Methylation States and Detects Sequence Motifs	89% AUC
CNNC	ANN	scRNA-seq	Predicting Transcription Factor Target	~70% Accuracy for Multiple Experiments
DanQ	CNN and RNN	DNA-seq	Predicting the Function of DNA Directly from Sequence alone	AUC score ~70%
FBGAN	GANs	DNA-seq	Optimize the Synthetic Gene Sequences	Train accuracy 0.94 test accuracy 0.84

TABLE 4. Difference between RNA-FM and PlantRNA-FM

Component	Transformer	BERT
Architecture Type	Encoder-Decoder: The model consists of an encoder to process input and a decoder to generate output. Commonly used in tasks like machine translation.	Encoder-only: The model focuses only on the encoder part of Transformer, making it more suitable for text-understanding tasks.
Attention Mechanism	Self-Attention + Multi-Head Attention: Each token attends every other token in the sequence to capture relationships, regardless of distance.	Self-Attention + Multi-Head Attention: The same mechanism as Transformer but applied bidirectionally to capture full contextual information.
Input Format	Full sequence input without masking: The input is processed as-is, without any part being hidden or modified.	Randomly masked input tokens (Masked Tokens): 15% of the input tokens are randomly replaced with a special [MASK] token, forcing the model to learn from the surrounding context.
Pretraining Tasks	None: The original Transformer model does not use a pretraining phase and is directly trained on specific tasks like translation.	Masked Language Modeling (MLM) + Next Sentence Prediction (NSP): Pretraining tasks that teach the model to predict missing words and understand sentence relationships, respectively.
Output Type	Generative tasks: The model generates new sequences, like translating a sentence from one language to another.	Understanding tasks: The model comprehends the input sequence to perform tasks like text classification, named entity recognition (NER), and question answering (QA).