Front. Digit. Educ., 2025, 2(1): 14
https://doi.org/10.1007/s44366-025-0051-1

REVIEW ARTICLE

AI-Empowered Genome Decoding: Applications of Large Language Models in Genomics

Shaopeng Li, Weiliang Fan, Yu Zhou

College of Life Sciences, Taikang Center for Life and Medical Sciences, RNA Institute, Wuhan University, Wuhan 430072, China

Abstract Large language models (LLMs) have transformed natural language processing with their improved performance compared with previous methods and have shown great potential to be adopted in other fields. The sequential nature of genomics data, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins, makes it akin to human natural language, supporting the application of LLMs. Currently, LLMs have only been applied to genomic research for about four years but have already achieved significant advances in many challenging and important problems. This review summarizes the recent progress of applying LLMs in genomic research, including developing biological foundation models for protein, DNA, and RNA, as well as specialized models for interaction prediction, single-cell analysis, and structure prediction. The review discusses the challenges and potentials of adopting new advancements in LLMs for genomic applications and proposes several practical projects for integrating LLMs into genomics teaching and learning.

Keywords large language models, deep learning, genomics, Wuhan University, higher education

1 Brief Overview of Large Language Models

Language models are machine learning models trained to analyze statistics and probabilities of words in phrases and sentences. In 2017, transformer architecture with the self-attention mechanism was proposed (Vaswani et al., 2017). Based on this work, an encoder-only transformer model, bidirectional encoder representation from transformers (BERT), was introduced by researchers at Google and achieved an unprecedented level over previous state-of-the-art methods in the natural language processing field (Devlin et al., 2019). Moreover, the two-step paradigm, pre-training and fine-tuning, is established, in which models are usually trained on an unlabeled large dataset and then trained on a task-specific small dataset for different downstream tasks. With the help of this innovative paradigm, many new models have been developed, including bidirectional and auto-regressive transformers (BART), generative pre-trained transformer (GPT) series products, and XLNet.

Researchers find that the performance of those pre-trained language models can be improved by scaling the model size and dataset volume (Kaplan et al., 2020). Moreover, it indicates that large-size language models can obtain more abilities in solving complex tasks without model architecture optimization (Wei et al., 2022). Therefore, language models with many parameters trained on a huge amount of data, termed LLMs, have received increasing attention and shown great potential in many areas (Shanahan, 2024).

These LLMs can be classified into four mainstream architectures, including encoder–decoder architecture, decoder-only architecture, encoder-only, and state space models (SSMs). First, the encoder–decoder architecture, the canonical transformer model, consists of two major components, an encoder and a decoder. The encoder takes the input sequence and generates latent space representations. The decoder takes the encoded input and predicts target sequences autoregressively. The BART and transferring text-to-text transformer (T5) are representative encoder–decoder models. Second, in the decoder-only architecture, input and output tokens are processed by stacked decoder layers incorporating masked multi-head attention blocks. Decoder-only architecture is well-suited for generation tasks, and the success of GPT demonstrated its capabilities for tasks like sequence generation and missing word prediction. Many LLMs

Received January 10, 2025; revised February 25, 2025; accepted March 8, 2025
Yu Zhou (✉)
E-mail: yu.zhou@whu.edu.cn

adopted this architecture, including GPT-series, bigscience large open-science open-access multilingual language model (BLOOM), open pre-trained transformers (OPT), pathways language model (PaLM), and LLM meta AI (LLaMA). Third, the encoder-only architecture only consists of stacked encoder layers and is not auto-regressive. This architecture provides better performance in classification tasks like sentimental analysis. BERT is an example of encoder-only architecture. Fourth, SSMs have been proposed to enhance efficiency and deal with long inputs. Comparing with conventional LLMs with self-attention mechanisms that have the quadratic computational complexity of $O(N^2D)$ in which N is token numbers and D is feature dimensions, SSMs can generate outputs recursively like recurrent neural networks (RNNs), which makes the decoding process faster for only one previous state should be visited. However, computational efficiency comes at the cost of inferior performance compared with transformer. Mamba, retentive networks, and Hyena are representative of SSMs architectures.

Many forms of biological data, including DNA, RNA, protein sequences, expression profiles, chemical compounds, and interaction networks, can be interpreted as a language-like symbolic framework in which vocabulary is combined with certain grammar and produces contextual meaning. LLMs are introduced to facilitate the understanding of these systems. Applying LLMs in genomics is growing rapidly, as Figure 1 summarizes some important foundation models by timeline.

2 Foundation Models for Genomic Research

2.1 | Generalized LLMs in Natural Language for Assisting Genomic Research

The human knowledge of the biological field is

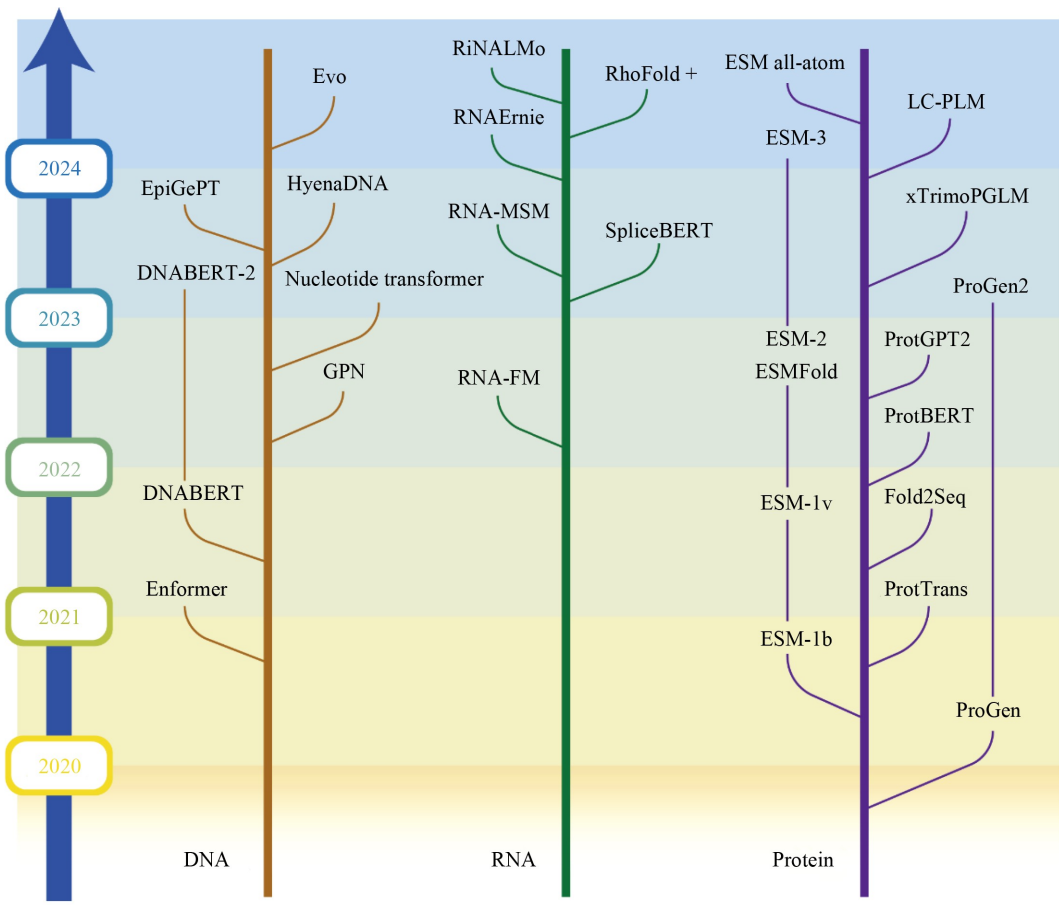

Figure 1 A brief timeline of major LLMs in genomics for DNA, RNA, and protein languages. DNA: deoxyribonucleic acid, RNA: ribonucleic acid, GPN: genomic pre-trained network, RiNALMo: ribonucleic acid language model, RNA-MSM: multiple sequence alignment-based RNA language model, BERT: bidirectional encoder representation from transformers, RNA-FM: RNA foundation model, ESM: evolutionary scale modelling, LC-PLM: long-context protein language model, ProtGPT: protein generative pre-trained transformer.

published through research papers, technical blogs, and textbooks. Natural language models trained as generalized dialogue models, such as GPT-4 (OpenAI et al., 2024) and LLaMA (Touvron et al., 2023), are based on a vast of available corpora, including this biological knowledge.

2.1.1 LLMs for Interactive Dialogue

Generalized foundation models have shown surprising capabilities in answering medical and biological questions with proper prompt engineering (Nori et al., 2023). The models outperform fine-tuned specific-purpose dialogue models like Med-PaLM2 in several tasks, like parsing research papers, synthesizing sophisticated knowledge, and making inferences in a given context (Singhal et al., 2023). Genomic researchers frequently encounter unfamiliar concepts in navigating genomic literature, such as new algorithms for researchers with genetics, or specific genes and pathways for researchers with a bioinformatics background. LLMs can reason across given reference papers, distil information about specific concepts into a refined description, and reduce the time needed for looking up references (Simon et al., 2024).

The main drawback of the LLMs as a research consultant is their tendency to make up false facts, such as referring to non-existent papers and generating persuading texts with factual errors. This is a well-known limitation in the generative language model, named hallucination (Ji et al., 2023). Multiple prompt-engineering approaches can be used to reduce hallucination by requiring the model to generate its step-by-step reasoning procedure (Kojima et al., 2023), take multiple reasoning paths (Wang et al., 2023), and perform self-correction with the external data source (Gou et al., 2024). Reducing model temperature and providing reference paper directly to the model can reduce hallucination. Even with these approaches, however, manual fact-checking is still needed for any information generated by the model.

2.1.2 LLMs for Code Generation

Natural language models have been proven to have great potential to write codes (Chen et al., 2021). Considering the vast range of bioinformatics software packages, it is time-consuming for a researcher to understand new software packages in genomics research, especially those restricted to field-specific usage. In a recent evaluation for bioinformatics programming (Tang et al., 2024), GPT-4 solved over 60% of Java programming problems and over 55% of Python programming problems in bioinformatics, showing that generalized natural language models owned knowledge of bioinformatics software and could invoke them correctly. With these capabilities, LLMs can be used to generate code snippets with annotations and code contexts, identify bugs in the programming, and generate potential fixes as suggestions (Majdoub & Charrada, 2024). LLMs are required to explain existing code to help researchers understand complex scripts and adapt to a new algorithm. These features of LLMs boost coding productivity and allow genomic researchers to focus more on interpreting the biological meaning of the data.

In the chemistry field, LLMs have been used to manipulate existing software and natural language interfaces (Bran et al., 2024). For genomics and bioinformatics, pre-prints are aiming at similar targets, such as BioMANIA (Dong et al., 2023). These tools are not widely used in genomics research but still suggest a promising approach to make complex genomics tools accessible to a broader scientific community.

2.1.3 Ethical Concerns of LLMs for Research Assistance

Some users might consider LLMs with interactive capability as creatures capable of thinking. Genomic researchers using LLMs for research assistance should be reminded that LLMs are network models trained on the objective of predicting the next token, which is different from humans and affects the biases in training datasets (Shanahan, 2024). The tendency to anthropomorphize LLMs is dangerous, as it leads to overestimation of the capability and reliability of language models, misconceptions of biological information, and academic misconduct.

There are two main threats to the vulnerability of LLMs, input attacks and training data extraction attacks. First, prompt injection is used to hijack the output of LLMs, in which hackers use malicious input to make the model disinform other users. Unlike chatbots powered by LLMs deployed to social media platforms, most LLM research assistants use independent context for each conversation. This difference limits the risk of prompt injection for LLMs used for research assistance. Second, training data extraction attacks are more dangerous for LLMs fine-tuned on biological data, especially for those fine-tuned on clinical data. In this attacking method, hackers use malicious input to make the model directly provide verbatim text sequences from training data. To address this issue, de-identification should be properly performed on the training dataset to avoid the threat fundamentally.

2.2 | Generalized LLMs for Biological Data Analysis

Compared with natural language models trained to

generate text sequences, biological language foundation models are trained with sequential biological data, such as sequences of biomolecules and serializable biological data. LLMs are trained to process different modalities and applied to a variety of fields in genomic research. The key difference between generalized foundation methods and former deep-learning-based methods is training strategy. The former deep-learning-based methods rely on supervised learning on large labelled datasets. While genomics research generates a huge amount of unlabeled data, foundation models use the masked language modelling approach and allow the model to learn the inherent grammar from vast amounts of unlabeled data and to be fine-tuned with fewer labelled data for specific tasks. These functions show a major advantage over former supervised learning approaches (Simon et al., 2024).

2.2.1 Foundation Models for Protein Sequences

Deep learning methods have been applied successfully in protein research, especially in protein structure and function prediction (Sapoval et al., 2022). Methods, like AlphaFold, have changed the paradigm of protein research (Jumper et al., 2021), and DeepGO has improved ontology classification compared with traditional methods, like finding similar sequences with known functions using the basic local alignment search tool (Kulmanov et al., 2018). However, most of those methods are based on multiple sequence alignment, and pre-trained language models have only been applied to build protein foundation models recently. Though most of these models are initially designed for specific purposes, they are proven to be reliable foundation models.

ProGen-series models are focused on protein sequence generation initially, but are widely used as foundation models for downstream tasks. ProGen has been pre-trained on 280 million raw protein sequences and fine-tuned with control tags for protein properties to generate unique proteins containing required sequences and tags (Madani et al., 2023). The model contains 1.2 billion trainable parameters. ProGen2 provides a model family for different infrastructures and increases the maximum scale to 6.4 billion parameters (Nijkamp et al., 2023). Aside from being fine-tuned to tasks like ProGen, ProGen2 are pre-trained on natural antibodies, called ProGen2 with an observed antibody space database (ProGen2-OAS) (Olsen et al., 2022). It is an antibody-specific base model and is used to design antibodies with protein aggregation propensity and solubility.

The evolutionary scale modelling (ESM) series, on the contrary, initially focused on structure prediction and proved to be a reliable foundation model for other downstream tasks. Compared with ProGen series models, ESM series models show more rapid growth in scale. ESM-1b is the first published ESM series model, which has about 650 million parameters and is pre-trained on 86 billion amino acids across 250 million protein sequences (Rives et al., 2021). ESM-2 provides a model family with different scales, with the largest model using 15 billion parameters and an improved dataset sampled across UniRef50 and UniRef90, and contains about 65 million protein sequences (Lin et al., 2023). ESM-3 further pushes the largest model scale to 98 billion parameters and is pre-trained on 2.78 billion nature protein sequences plus 0.37 billion synthetic sequences (Hayes et al., 2025).

ESM models were initially applied to structure prediction. As language models learn the evolutionary patterns in similar proteins, these models in the ESM series do not need multiple sequence alignment (MSA) data for accurate predictions. ESMFold, a model based on ESM-2 fine-tuned for structure prediction on sequences with MSA, is 60 times faster compared with AlphaFold2 with minor accuracy loss. On sequences without MSA, ESMFold shows superior accuracy. These features make it a promising tool in genomic research, especially in metagenomic research, with many unknown proteins with minimal known structures as reference. ESMFold is used to construct the ESM metagenomic atlas and provides structure prediction for 617 million metagenomic proteins. Moreover, the ESM model series also serves as a foundation model for other downstream tasks. BioLLMNet uses ESM-2 for RNA-protein interaction prediction (Tahmid et al., 2024a), and GLM uses ESM-2 for protein co-regulation prediction (Hwang et al., 2024). There are also approaches to fine-tune ESM-2 to multiple downstream tasks, such as protein stability and subcellular location (Schmirler et al., 2024).

2.2.2 Foundation Models for DNA Sequences

DNA and RNA have lower information density than protein, as nucleotide sequences are four-letter encoded compared with 20-letter encoded protein sequences. This is a main challenge in building nucleotide foundation models, requiring a larger network scale and more complexity. Compared with protein models, the development of DNA and RNA foundation models has just recently started.

Like protein foundation models, DNA foundation models (DNA-FMs) are designed to overcome the limitations of former deep learning models, such as the limitation of annotated data and low efficiency in adapting to different tasks by transfer learning. In this review, the nucleotide transformer, a recent foundation model pre-trained on DNA sequences, is taken as an example (Dalla-Torre et al., 2024).

The nucleotide transformer is constructed with the transformer architecture inspired by LLMs like BERT and GPT, with nucleotide sequences as sentences and 6-mer motifs as words. This architecture allows the model to process DNA sequences by considering the context of each nucleotide. Like other foundation models, this model is pre-trained by learning to predict masked nucleotides within a sequence, forcing it to understand the underlying grammar and context of genomic information.

Four language models with different model scales and training datasets are built to understand how these differences can influence the models' ability to learn generalizable genomic information. Generally, models with larger parameter scales and larger training datasets will outperform smaller models in most downstream tasks. The nucleotide transformer 2.5b multi-species model is the largest model of the four models and has outperformed other models in most tasks. Moreover, the model trained on sequences from multispecies has outperformed models trained on sequences only from the human genome, even on specific tasks on the human genome. This suggests that the model can learn from the evolutionary diversity across different species and allow it to identify and focus on conserved fundamental features for their functional importance. Generally, DNA-FMs can benefit from larger parameter scales and more diverse training datasets, similar to natural language models. These FMs show great potential for LLM applications in genomics research.

The nucleotide transformer has been fine-tuned and tested on different downstream tasks for genomic research. It shows strong performance in molecular phenotype prediction, such as splice site prediction, identification of promoters and enhancers, and characterization of histone modifications. This model can achieve high accuracy with few labelled data. Adding data from other species helps to increase accuracy and makes the FM valuable for less-characterized genomic features and non-model organisms. It has been demonstrated to be a powerful tool for the prediction of genetic variant effects. The model matches and outperforms the existing supervised models on this task, shows a great advance in zero-shot samples, and has the potential for application in unannotated organisms. The nucleotide transformer's attention head shows good alignment with important genomic elements, such as transcription factor binding sites, and suggests that the attention map can be further parsed to discover novel regulatory elements or find uncharacterized known regulatory elements.

Evo is a recent DNA-FM using a different approach (Nguyen et al., 2024). The model is based on a new architecture called Hyena, a state-space model using long convolution filters instead of transformers (Poli et al., 2023). As the convolution filters are much faster than transformers to process long token context, Evo can extend the context length to 131,072 tokens, compared with 1,000 tokens for the nucleotide transformer. The extended context range allows Evo to discover long-range interactions in the genome and extract deeper information from the sequence. Evo designs DNA sequences on a genomic scale and is fine-tuned to zero-shot tasks, like predicting mutational effects on protein and non-coding RNA functions.

2.2.3 Foundation Model for RNA Sequences

Building an RNA-FM is more complex than building a DNA model. RNAs are flexible molecules with numerous possible secondary and tertiary structures, and different RNA types have different interaction mechanisms with other molecules (Holbrook, 2005). An early approach in 2022 has proposed an RNA-FM, a BERT-based RNA-FM pre-trained on 23 million non-coding RNA sequences (Chen et al., 2022). Though this RNA-FM performed well on non-coding RNA classification, its performance in some downstream tasks, like RNA–RNA interaction and RNA secondary structure prediction, has not been improved compared to former deep learning methods. There are also several models like RNA–BERT using the BERT model as an RNA embedding approach for downstream tasks (Akiyama & Sakakibara, 2022). In 2024, RNAErnie was introduced with motif-aware pre-training and type-guided fine-tuning (Wang et al., 2024a). This review takes RNAErnie as an example for introduction.

For motif awareness in pretraining, RNAErnie is based on enhanced representation through the knowledge integration (ERNIE) model (Sun et al., 2019). A major difference between ERNIE and BERT is the masking strategy. BERT masks each token with the same probability, while ERNIE uses prior knowledge to propose a multi-level masking strategy. In the first learning stage, ERNIE uses a similar masking method to BERT. In the second learning stage, ERNIE masks continuous phrases as small groups of words together as a conceptual unit, instead of single tokens. In the third learning stage, ERNIE masks continuous entities in which words contain an abstract concept and an existing entity, shorter than phrases, instead of single tokens. In this way, the model can learn richer semantic information. RNAErnie uses this feature to extract knowledge from the subsequence and motif levels. In the base learning stage, 15% of the bases are masked randomly for unsupervised learning. In the subsequence learning stage, the bases are masked randomly and continuously with lengths ranging from 4 base pairs (bp) to 8 bp. In the motif stage, motifs from ATtRACT and SpliceAid (Giudice et al., 2016; Piva

et al., 2012), combined with some of the most frequent motifs from the whole dataset, are recognized randomly and masked continuously. With this pre-training strategy, RNAErnie can learn context-related functions of nucleobases without direct usage of the known motifs in embedding.

For type-guidance in fine-tuning, RNAErnie uses a stacking strategy. RNAErnie is trained to predict the possible RNA type, like mRNA and lnRNA, as a special ending token. In fine-tuning for downstream tasks, the downstream head has different sub-networks for each RNA type. These sub-networks share some parameters for type-unrelated RNA information. The modes predict the top-k possible RNA types which means the top k most probable options, and k sub-networks take parallel inputs simultaneously. The outputs of these sub-networks are further ensembled with the RNA type possibility as output weight. With this fine-tuning strategy, RNAErnie can adapt to the distributional differences and function differences between different RNA types, which helps downstream tasks.

RNAErnie has been adapted to RNA sequence classification, RNA–RNA interaction prediction, and RNA secondary structure prediction. In all these downstream tasks, RNAErnie surpasses existing methods, including former foundation models like RNA-FM, which proves its ability to learn robust and generalizable representations of RNA sequences.

Recently, a preprint has proposed DGRNA, a new RNA-FM trained on a larger set with 100 million RNA sequences compared to 23 million used by RNAErnie and a state-space-based language model Mamba-2 (Yuan et al., 2024). Compared with RNAErnie, Yuan et al. (2024) showed major improvements in various downstream tasks without specific adaption to RNA features. The ongoing development of LLMs has great potential to further enhance biological foundation models.

2.2.4 FMs for Specific Research Fields

Though FMs can be enhanced by the larger scale and diversity in pre-training datasets, some research fields still need task-specific datasets for model training for better results. For example, the language model for decoding untranslated regions (UTR-LM) is an FM specifically pre-trained on 5′ untranslated region (UTR) sequences from multiple species (Chu et al., 2024). This model targets the regulatory function of the 5′ UTR and is fine-tuned to multiple regulation-related downstream tasks, such as predicting ribosome loading, mRNA expression level, and internal ribosome entry site. The UTR-LM performs well in these downstream tasks and outperforms generalized models like RNA-FM. For RNA splicing, SpliceBERT is unsupervised pre-trained on primary RNA sequences and uses the masking training approach similar to other BERT-based models. The extracted hidden state from encoder layers shows a clear distinction between conserved and non-conserved sites. Moreover, FM can capture evolutionary conservation. SpliceBert is further fine-tuned to predict RNA splicing site prediction and shows improvements compared to the RNA-FM (Chen et al., 2024a).

There are also works using methods like generalized models but focusing on a specific downstream task. The genomic pre-trained network (GPN) is unsupervised pre-trained on genomic sequences, similar to other DNA-FMs. The unsupervised clustering of contextual embedding has shown the ability to predict different genomic regions through the sequence and indicates that the FM has learned the structure of the genome (Benegas et al., 2023). However, GPN focuses on variant effect prediction, and there is a major difference between GPN and other generalized models. GPN uses a BERT-like structure, but has replaced transformer blocks with convolutional blocks and is trained on a smaller dataset for faster training speed. Training weights of repeat sequences are optimized carefully for the model to achieve better performance in variant effect prediction.

Currently, FMs trained with specialized training datasets and generalized datasets are both advancing rapidly. Specialized models often show improvements in specific downstream tasks, but in most cases, the improvements are minor. For example, SpliceBERT achieves an F1 score of 0.957 in zebrafish splice site prediction, while RNA-FM, a generalized model focusing more on non-coding RNA, achieves 0.937. A recent evaluation of DNA language models also shows that FMs specifically trained on the human genome have no significant advantages compared with models trained on benchmarks of mammalian genomic data and multi-species genomes, including bacteria and fungi, in performance evaluation on benchmarks of mammalian genomic data (Patel et al., 2024). Generalized models, generally, benefit from a larger and more diverse dataset for pre-training. For sequences with specific biological regulation pathways, specialized models outperform generalized models with fewer computational requirements.

3 Other Applications in Genomic Analysis

Besides tasks with sequential inputs that can be adopted with fine-tuning of FMs, there are other applications requiring further modifications to the model structure. Interaction prediction needs multi-modal capability,

while single-cell omics data are not sequential natively. Some new approaches in structure prediction combine the language model and diffusion model.

3.1 | Specialized Model for RNA–RNA and Protein–RNA Interaction Prediction

The interactions between RNA and other biomolecules play important roles in regulating different biological procedures. Multiple techniques have been used to decipher such interactions, like RNA *in situ* conformation sequencing (RIC-seq) for RNA–RNA interaction (Ye et al., 2024) and cross-linking and immunoprecipitation coupled with sequencing for RNA–protein interaction (Lin & Miles, 2019). With the vast volume of omics data generated by these techniques, there are approaches with deep learning methods to predict these interactions and reveal hidden interaction patterns. Some of the approaches used a simple convolution neural network (CNN) (Pan & Shen, 2018), while others used a combination of CNN and recurrent networks (Grønning et al., 2020). Interaction prediction of RNA with other biomolecules is a challenging task for language models, as most FMs are pre-trained to understand the biological information in a single sequence.

This review takes BioLLMNet as an example of applying LLMs to RNA interaction prediction (Tahmid et al., 2024a). BioLLMNet utilizes a multimodal approach by using different language model encoders for different biomolecules. In the BioLLMNet architecture, RNA sequences are encoded by BiRNA-BERT (Tahmid et al., 2024b), protein sequences are encoded by ESM-2 (Lin et al., 2023), and small molecules are encoded by Mole-BERT (Xia et al., 2023). For each interaction combination, the output of the RNA language model is transformed with a multi-layer perceptron (MLP) to align the feature space with the other language model's output, and a gated weight MLP is used to perform a weighted average of two aligned embeddings, followed by a multi-layer MLP as a final prediction head to predict the interaction. By using LLMs as embedding encoders, BioLLMNet can avoid using predefined features, like sequence motifs, and capture more information with pre-trained foundation models. With this strategy, BioLLMNet combines multi-modal features and outperforms former deep-learning-based methods on interaction prediction tasks.

3.2 | Specialized Model for Single-Cell Data Analyses

Single-cell omics is an important approach in genomics research. This approach reveals cell-specific features obscured in bulk sequencing, crucial for cell fate determination, cell–cell interaction, immunology, and oncology. However, single-cell omics data are a unique data type compared with DNA, RNA, and protein sequences, as these data are not natively sequential. To address this issue and utilize the language model on single-cell data processing, researchers use different approaches, such as converting gene expression levels to a sorted gene list in text format or using the embedding of expressing genes as words.

Cell type annotation is one of the most important tasks in single-cell data processing. Traditional methods often rely on the expression level of predefined marker genes which struggle with technical noise, batch effects, and sparsity (Tung et al., 2017), and are not effective for poorly characterized cells (Clarke et al., 2021).

Smaller transformer models are initially used on cell annotation tasks. TransCluster and CIForm apply supervised training on annotated scRNA-seq data and show superior performance compared with traditional methods (Song et al., 2022; Xu et al., 2023a). Moreover, scTransSort applies a self-attention model trained with unlabeled gene embedding for cell type annotation and reduces the need for manually labelled features or reference sequencing data (Jiao et al., 2023). The transformer for one-stop interpretable cell-type annotation introduces an interpretable approach using a multi-head self-attention model, annotates cell types using biological pathways and regulations, and provides insights into the biological character of different cell types (Chen et al., 2023).

Several LLMs have promising results on this task. The single-cell BERT applies the BERT model pre-trained on a massive single-cell RNA sequencing dataset in an unsupervised approach to learning general gene-gene interactions, and transfers to the cell type annotation task with specific single-cell RNA sequencing data in a supervised approach (Yang et al., 2022). Recently, larger models trained on human language have also been used on cell-type annotation tasks. GPTCelltype used GPT-4, a general-purpose LLM on cell type annotation (Hou & Ji, 2024). The model converts the gene expression list and marker gene list into readable text format and uses the language model directly for cell type annotation. Though the output of GPTCelltype is still affected by AI hallucination and human involvement is needed for prompt engineering, GPTCelltype still reveals that general-purpose language models trained on human language can handle biological data correctly without transfer training and potentially reduce the need for specialized expertise and re-training cost.

Beyond cell type annotation, LLMs are also used for downstream analysis. The ScRAT aims at the challenge of connecting single-cell RNA sequencing data with phenotype differences, such as different

diseases (Mao et al., 2024). A multi-head attention encoder network is used to aggregate sample cell embeddings from each sample and identify crucial cells for each phenotype without predefined annotation. This method enables researchers to uncover novel relationships between specific cell transcriptomic status and phenotypes. STGRNS is designed to infer gene regulatory networks from single-cell RNA sequencing data (Xu et al., 2023b). STGRNS uses a gene expression motif technique to convert gene pairs into contiguous sub-vectors for transformer encoder input, improve inference accuracy, and outperform traditional methods. There are also algorithms for other tasks, such as iSEEEK for multi-dataset integration (Shen et al., 2022). Geneformer is a context-aware attention-based model pre-trained on single-cell RNA sequencing data for fundamental purposes (Theodoris et al., 2023). With further transfer learning, Geneformer can be fine-tuned towards multiple tasks with limited task-specific data.

There are two approaches are applied to enhance further the potential of LLMs in single-cell data processing. The first approach is to train a fundamental multi-purpose language model on single-cell data to decipher the cells' language. The cell pre-trained language model (cellPLM) is an early approach to treat genes as tokens and cells as sentences and outperforms former models in diverse downstream tasks (Wen et al., 2023). In 2024, single-cell GPT (scGPT) was published as a generative pre-trained transformer model, was trained on a larger dataset of over 33 million cells, and could be optimized for multiple downstream tasks (Cui et al., 2024). Recently, the single-cell foundation (scFoundation), an even larger foundation model with over 100 million parameters trained on over 50 million single-cell RNA sequencing profiles covering about 20,000 genes (Hao et al., 2024). The model achieves state-of-the-art performance in multiple downstream analysis tasks, including cell type annotation, gene relationship, drug response, and gene module inference.

The second approach is to "translate" biological information to text embedding and improve the performance of foundation models trained with texts. Cell2Sentence (C2S) is a method to translate gene expression data into cell sentences and provides a method to adapt any language models to a single-cell context directly (Levine et al., 2023).

The C2S generates valid gene expression levels based on input cell types and predicts accurate cell types from given expression levels. On the contrary, GenePT parsed the description texts with a language from the National Center for Biotechnology Information to generate text embedding for genes, and composed gene embedding by expression level for a cell embedding (Chen & Zou, 2023). With a much simpler approach, GenePT achieves similar performance on tasks like gene classification and cell type annotation compared with models trained on biological datasets. The combination of two approaches, like single-cell Mulan, represents cells as structured cell sentences, further trains the language model on single-cell RNA sequencing data translated into sentences, and performs different downstream tasks guided by task-specific prompts (Chen et al., 2024a).

3.3 | Specialized Model for Biomolecular Structure Prediction

Biomolecular structure prediction has been advanced by applying deep learning methods. Tools, like AlphaFold and RoseTTAFold, have achieved near-experimental accuracy in protein structure prediction, and methods, like MXfold2 and SPOT-RNA2, show significant improvements in prediction accuracy over traditional methods (Baek et al., 2021; Jumper et al., 2021; Sato et al., 2021; Singh et al., 2021).

LLMs are recently applied in biomolecular structure prediction. Some early approaches utilize a transformer attention map unsupervised trained on protein sequences to predict protein contacts, which shows that the language model can learn structural information (Rao et al., 2020). Rivers et al. (2021) scaled language models to 86 billion amino acids across 250 million protein sequences, which allowed models to learn further sequence representation and contained information about secondary structure prediction as well as mutational effect prediction. In 2022, larger models combining the Transformer-XL architecture and T5 language encoder model were applied to protein sequences. ProtTrans surpassed state-of-the-art methods without requiring multiple sequence alignment data in secondary structure prediction and implied that large language models had a high potential for the language of protein (Elnaggar et al., 2022). To further apply an LLM to atomic-resolution 3D structure prediction, ESMFold brings a sequence-to-structure predictor based on an LLM scaled up to 15 billion parameters. Compared to existing alignment-based methods, ESMFold achieved comparable accuracy with significantly higher computation speed and enabled structure predictions on more than 600 million metagenomic proteins (Lin et al., 2023).

Applying LLMs to nucleotide acid structure is a more recent but developing rapidly field. Early language models, like BERT, have been used to create RNA FMs like RNA-FM (Chen et al., 2022), but its performance in structure prediction shows limitations compared with smaller models like SPOT-RNA2 tested in multiple sequence alignment-based RNA language model (RNA-MSM) paper. The main challenge compared to proteins is the lower information density of four-letter coded DNA and RNA sequences compared with 20-letter

codec protein sequences. To address this issue, RNA-MSM used an unsupervised multiple sequence alignment-based approach to find homologous sequences in less conserved RNA sequences (Zhang et al., 2024). The attention maps and embeddings from RNA-MSM contain significant structural information, which allows the model to be fine-tuned for downstream tasks like predicting 2D base pairing probabilities and 1D solvent accessibilities, and achieve superior performance compared with existing state-of-the-art methods. As another approach, RNAformer designed a scalable axial-attention-based model to predict secondary structure from RNA sequence directly (Franke et al., 2024). RNAformer employed a homology-aware data pipeline to overcome the homology bias in training data and testing data and achieves state-of-the-art performance on RNA secondary structure prediction without the need for any reference sequence.

Another method to overcome the difficulties in RNA-FMs is to train larger models over larger datasets. RiNALMo is a ribonucleic acid language model recently published on arXiv and has trained 650 million parameters over 36 million non-coding RNA sequences from multiple datasets. RiNALMo overcame the limitations of existing deep learning methods in generalizing unseen RNA families in RNA secondary structure prediction and suggested a promising future for LLMs in revealing hidden information within nucleotide acid sequences. Another recent preprint method, RhoFold +, demonstrated using a LLM for single-chain RNA 3D structure prediction (Shen et al., 2024). RhoFold + offered a fully automated end-to-end pipeline and showed superior performance over existing methods.

Some recent approaches, like AlphaFold3 and RoseTTAFold all-atom, utilized the diffusion model to increase performance on complex tasks, such as biomolecular interaction and small molecule binding (Abramson et al., 2024; Krishna et al., 2024). Diffusion models contained transformer blocks for prompt processing but were typically pre-trained to denoise data iteratively, unlike language models pre-trained to predict the next token in a sequence. There were also approaches to combine the diffusion model with the language model, such as diffusion protein language model-2 (DPLM-2) (Wang et al., 2024b). DPLM-2 employed a multimodal diffusion protein language model, combined structure and sequence information in the same backbone model, used separated tokenizer and output heads for structure and sequence, and thus enabled multimodal applications, such as predicting sequence from structure, protein function embedding, and designing protein with specific structure and sequence.

4 Discussion

The review represents a brief review of the expanding usage of LLMs in genomics research, including commonly used biological FMs like ESM-2 and RNA-FM, specific-target models like GPN, and adaptors for natural language models to handle biological data like C2S.

4.1 | Summary of Representative FMs

As shown in Table 1, the review summarized representative FMs on their architectures, parameters, and unique features. The advantages and disadvantages of different architectures could be case-specific depending on different training strategies and data engineering approaches.

4.2 | Data Augmentation Strategy of LLMs in Genomics

For most FMs, the simple strategy to pre-train language models on unlabeled sequence data and then fine-tune them on specific tasks already works well. This strategy is applied broadly in most natural language models. Data augmentation is a commonly used training strategy in neural network models. For biological language models, random mutation is one of the most commonly used methods in genomic data. This approach selects some tokens from the input sequence dataset randomly and replaces them with selected tokens. For example, RNAErnie uses this data augmentation following a former approach used in UFold (Fu et al., 2022), which mutates randomly 20%–30% of nucleotides in RNA and uses Contrafold to generate ground truth labels for these sequences (Do et al., 2006).

However, LLMs generally use simpler augmentation approaches compared with former smaller natural language processing models which apply a combination of paraphrasing, noising, and sampling to generate augmented data with high diversity (Li et al., 2022). While GPT-3, a representative LLM, only used corpora replacements, LLaMA does not use data augmentation. Similarly, biological language models also used fewer augmentation approaches, the nucleotide transformer only uses random augmentation in smaller datasets and skips augmentation on the 1,000 genome dataset, as the added noise has a different pattern from the natural mutation in the genome. A possible hypothesis for this phenomenon is the pre-training strategy. Pre-training of LLMs does not require labelled data and allows the vast dataset of human

Table 1 Comparison of representative FMs

Model type	Model	Architecture	Parameters of the FMs	Features
DNA models	DNABERT (2021)	Encoder-only transformer	89 million	Early application of LLM on DNA sequence
	DNABERT-2 (2024)	Encoder-only transformer	117 million	Lower resource requirements compared with nucleotide transformer
	Nucleotide transformer (2024)	Encoder-only transformer	2.5 billion	Better performance with larger model and multi-species data
	HyenaDNA (2023)	Decoder-only Hyena	1.6 million	Similar performance with nucleotide transformer using fewer parameters
	Evo (2024)	Decoder-only striped Hyena	7 billion	Extended context length, up to 131 kilobytes and improved performance in sequence generation significantly
	Evo2 (2025)	Decoder-only striped Hyena2	40 billion	Further extended context length, up to 1 million base pair
RNA models	RNA-FM (2022)	Encoder-only transformer	23 million	Mainly focusing on the 3D structure of non-coding RNA and adapting to other downstream tasks
	SpliceBERT (2024)	Encoder-only transformer	19.4 million	Focusing on RNA splicing
	RNA-MSM (2024)	Encoder-only transformer	96.5 million	Based on aligning homologous sequences and adapting to multiple downstream tasks
	RNAErnie (2024)	Encoder-only transformer	105 million	Generalized RNA FM for different types of RNA and adapted well to different downstream tasks
	RiNALMo (2024)	Encoder-only transformer	650 million	Focusing on non-coding RNA
	RhoFold + (2024)	Based on RNA-FM	/	RNA 3D structure prediction
Protein models	ESM-1b (2020)	Encoder-only transformer	650 million	Mainly focusing on learning evolution information for protein generation
	ESM-2 (2023)	Encoder-only transformer	15 billion	Improved in architecture and scale and better performed on multiple tasks
	ESM-3 (2025)	Encoder-only transformer	98 billion	Enabled guided protein generation with user prompts
	ProGen (2023)	Decoder-only transformer	1.2 billion	Focusing on generating homologous proteins of known protein family
	ProGen2 (2023)	Decoder-only transformer	6.4 billion	Increased model scale and better performance in novel protein generation and fitness prediction
	ProtBERT (2022)	Encoder-only transformer	16 million	Generalized protein foundation model requiring minimal resources
	ProtGPT2 (2022)	Decoder-only transformer	738 million	Mainly focusing on de novo protein generation

Notes. DNA: deoxyribonucleic acid, BERT: bidirectional encoder representation from transformers, LLM: large language model, RNA-FM: RNA foundation model, RNA-MSM: multiple sequence alignment-based RNA language model, RiNALMo: riboNucleic acid language model, ESM: evolutionary scale modelling, ProtGPT: protein generative pre-trained transformer.

language or biological sequencing data to be used as the training set. LLaMA uses a dataset containing 1.4 trillion tokens, and most of the training data are only utilized once during training. While smaller models often apply the same training data in supervised learning for hundreds of epochs. Large and diverged datasets might have decreased the importance of data augmentation.

Evo2, one of the most recent FMs, uses data augmentation in another way (Brixi et al., 2025). Instead of increasing diversity through random mutations, Evo2 uses data augmentation to attach extra information to sequences, like adding additional starting sequences to mRNA and adding gap sequences around exons. The additional information helps Evo2 to improve performance compared to former models and reveals a new strategy of data engineering for biological language models.

4.3 | Challenges and Potentials for LLMs in Genomics

A major challenge of biological LLMs is the intrinsic difference between biological data and human language, which acquires great achievements in natural language models inaccessible to biological language models. The latest progress in natural language models, like DeepSeek-R1, focuses on reasoning and reinforcing

learning and leads to a great improvement compared with former models (DeepSeek-AI et al., 2025). The reasoning feature of LLMs uses prompt engineering to let the model generate a thinking procedure in the natural language before generating the final result, and the contexts generated in this procedure effectively increase the quality of the result. Reinforcement learning uses existing exams and leverages prompt engineering to let the model generate formatted answers. As natural language is used in raising questions, thinking, and answering with easily accessible training data, this strategy is very effective in natural language models. However, transferring thinking progress and reinforcement learning to the biological language model can be a challenging task. The thinking procedure of the biological data sequence is unclear and may not be describable with the biological sequence itself. Reinforcement learning on biological data is also affected by the lower data quality compared with exam questions. As a result, recent progress in LLMs can be natural-language specific and is not effectively transferred to biological data.

Improvements in biological data encoding and the development of biological data-specific prompt-engineering methods help to address this problem. Currently, FMs for single-cell data analysis focus on this area early, as single-cell data is not originally sequential. In the future, data encoding and prompt engineering will be crucial in biological LLMs for other fields.

Another major limitation of LLMs is their rapid growth on scale, which results in the growth of computational resource requirements. Models with billions of parameters are untrainable for most researchers and limit the progression of FM pre-training to a few corporations. Fine-tuned models are unaffordable for many laboratories. Most researchers perform inference with large models or use the results generated by these large models. However, for some specific downstream tasks like perturbation response (Wenteler et al., 2024), models need to be trained or fine-tuned on specific high-quality datasets to capture a stable understanding of biological information. Some recent research articles address this limitation by releasing multiple models with different sizes and allow researchers with accessible computational resources to fine-tune the model for specific purposes.

4.4 | Integration of LLMs into the Genomics Course

Integrating LLMs into the genomics course facilitates students to better understand and apply genomic knowledge. Introducing the basics and advances of LLMs related to genomics in classroom teaching is encouraged. More importantly, the review proposes eight practical projects using LLMs to enhance the learning experience in genomics, including sequence generation, gene annotation tasks, protein sequence tasks, protein mutation effect prediction, protein structure tasks, RNA structure prediction, single-cell tasks, and epigenomics tasks.

First, for sequence generation, the student's experiment is arranged using DNA language models to generate a group of DNA sequences from a given prompt initial sequence and then cluster specific motifs from the generated sequences. The web service provided by the Arc Institute is used in this project, which provides the Evo DNA model for DNA sequence generating and scoring.

Second, for the gene annotation tasks, the student's experiment is arranged by deploying a DNA language model, annotating a genomic sequence, and comparing the annotation with Ensembl gene annotation. DNABERT is a model suitable for this task, as it is relatively small and deployable on most computers.

Third, for protein sequence tasks, the student's experiment is arranged to utilize a protein language model to generate embeddings for a group of protein sequences from different protein families and then cluster these embeddings for protein classification. The application programming interface (API) of the EMS3 protein model provided by Facebook is suitable for this task.

Fourth, for protein mutation effect prediction, the student's experiment is arranged by adding random mutation to different positions of the protein and then measuring the distance of embedding of a protein language model from the original protein to the mutated one. This experiment aims at gaining deep insights into the importance of each amino acid to the overall protein function using the ESMC API mentioned above.

Fifth, for protein structure tasks, the student's experiment is arranged by predicting protein structure with AlphaFold3 and ESMC using their public web service, followed by comparing predictions with the real structure. A further experiment is arranged by measuring the deviation of the two methods in different protein types.

Sixth, for RNA structure prediction, the student's experiment is arranged using an RNA language model to predict the 3D structures of a group of RNA sequences with known structures. The predictions will be compared with the ground truth to find the differences for different types of RNAs. RhoFold +, based on RNA-FM, is a smaller and deployable model for this experiment.

Seventh, for single-cell tasks, the student's experiment applies single-cell GPT data analysis. For this task, the Superbio team has provided single-cell GPT as three web applications, including reference mapping, cell annotation fine-tuning, and gene

regulatory network inference. The student experiment is arranged to deploy and fine-tune single-cell GPT locally. However, it requires a modern graphic processing unit to run effectively.

Eighth, for epigenomics tasks, the student's experiment utilizes an LLM to find methylation-related sites and a specific motif related to DNA methylation. The Mulan-Methyl model is provided by the University of Tubingen as a web service for finding methylation-related sites.

An increasing number of LLMs will be developed in genomics, and utilizing these LLMs in the classroom can help students better to learn genomics while also helping them to recognize the potential and limitations of existing tools.

Acknowledgments This study was supported by the Ministry of Science and Technology of the People's Republic of China (Grant No. 2023YFC2307802), the National Natural Science Foundation of China (Grant No. 82341023), and the Fundamental Research Funds for the Central Universities (Grant No. 2042022dx0003).

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statements The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data was anonymized before the statistical analyses were done.

Data Availability Statements The authors confirm that all data generated or analyzed during this study are included in this published article.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., & et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500.

Akiyama, M., & Sakakibara, Y. (2022). Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genomics and Bioinformatics*, *4*(1), Article 1.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., & et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876.

Benegas, G., Batra, S. S., & Song, Y. S. (2023). DNA language models are powerful predictors of genome-wide variant effects. *Biophysics and Computational Biology*, *120*(44), e2311219120.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting LLMs with chemistry tools. *Nature Machine Intelligence*, *6*(5), 525–535.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., & et al. (2025). *Genome modelling and design across all domains of life with Evo 2*. Arc Institute Manuscripts.

Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., Shen, T., & et al. (2022). Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv Preprint*, arXiv:2204.00300.

Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., & Han, J.-D. J. (2023). Transformer for one stop interpretable cell type annotation. *Nature Communications*, *14*(1), 223.

Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., & Yang, Y. (2024a). Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. *Briefings in Bioinformatics*, *25*(3), bbae163.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., & et al. (2021). Evaluating LLMs trained on code. *arXiv Preprint*, arXiv:2107.03374.

Chen, Y., Bian, H., Wei, L., Jia, J., Dong, X., Li, Y., Zhao, Y., Wu, X., Li, C., Luo, E., & et al. (2024b). Toward mastering the cell language by learning to generate. *bioRxiv Preprint*, bioRxiv:2024.01.25.577152.

Chen, Y., & Zou, J. (2023). GenePT: A simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv Preprint*, bioRxiv:2023.10.16.562533.

Chu, Y., Yu, D., Li, Y., Huang, K., Shen, Y., Cong, L., Zhang, J., & Wang, M. (2024). A 5′ UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence*, *6*(4), 449–460.

Clarke, Z. A., Andrews, T. S., Atif, J., Pouyabahar, D., Innes, B. T., MacParland, S. A., & Bader, G. D. (2021). Tutorial: Guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols*, *16*(6), 2749–2764.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, *21*(8), 1470–1480.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., & et al. (2024). Nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, *22*, 287–297.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., & et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning (version 1). *arXiv Preprint*, arXiv:2501.12948.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language

understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis: Association for Computational Linguistics, 4171–4186.

Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, *22*(14), e90–e98.

Dong, Z., Zhong, V., & Lu, Y. Y. (2023). BioMANIA: Simplifying bioinformatics data analysis through conversation. *bioRxiv Preprint*, bioRxiv: 10.29.564479.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., & et al. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127.

Franke, J. K. H., Runge, F., Köksal, R., Matus, D., Backofen, R., & Hutter, F. (2024). RNAformer: A simple yet effective model for homology-aware RNA secondary structure prediction. *bioRxiv Preprint*, bioRxiv:2024.02.12.579881.

Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2022). UFold: Fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research*, *50*(3), e14–e14.

Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT—A database of RNA-binding proteins and associated motifs. *Database*, *2016*, baw035.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: LLMs can self-correct with tool-interactive critiquing. *arXiv Preprint*, arXiv:2305.11738.

Grønning, A. G. B., Doktor, T. K., Larsen, S. J., Petersen, U. S. S., Holm, L. L., Bruun, G. H., Hansen, M. B., Hartung, A.-M., Baumbach, J., & Andresen, B. S. (2020). DeepCLIP: Predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic Acids Research*, *48*(13), 7099–7118.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., & Song, L. (2024). Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, *21*(8), 1481–1491.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., & et al. (2025). Simulating 500 million years of evolution with a language model. *Science*, *387*(6736), 850–858.

Holbrook, S. R. (2005). RNA structure: The long and the short of it. *Current Opinion in Structural Biology*, *15*(3), 302–308.

Hou, W., & Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods*, *21*(8), 1462–1465.

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., & Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. *Nature Communications*, *15*(1), 2880.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 248.

Jiao, L., Wang, G., Dai, H., Li, X., Wang, S., & Song, T. (2023). scTransSort: Transformers for intelligent annotation of cell types by gene embeddings. *Biomolecules*, *13*(4), 611.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., & et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv Preprint*, arXiv:2001.08361.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). LLMs are zero-shot reasoners. *arXiv Preprint*, arXiv:2205.11916.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., & et al. (2024). Generalized biomolecular modelling and design with RoseTTAFold All-Atom. *Science*, *384*(6693), eadl2528.

Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, *34*(4), 660–668.

Levine, D., Rizvi, S. A., Lévy, S., Pallikkavaliyaveetil, N., Zhang, D., Chen, X., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., & et al. (2023). Cell2Sentence: Teaching LLMs the language of biology. *bioRxiv Preprint*, bioRxiv:2023.09.11.557287.

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71–90.

Lin, C., & Miles, W. O. (2019). Beyond CLIP: Advances and opportunities to measure RBP–RNA and RNA–RNA interactions. *Nucleic Acids Research*, *47*(11), 5490–5501.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., & et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., & et al. (2023). LLMs generate functional protein sequences across diverse families. *Nature Biotechnology*, *41*(8), 1099–1106.

Majdoub, Y., & Charrada, E. B. (2024). Debugging with open-source LLMs: An evaluation. In: *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. New York: ACM, 510–516.

Mao, Y., Lin, Y.-Y., Wong, N. K. Y., Volik, S., Sar, F., Collins, C., & Ester, M. (2024). Phenotype prediction from single-cell RNA-seq data using attention-based neural networks. *Bioinformatics*, *40*(2), btae067.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., & et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. *Science*, *386*(6723), eado9336.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., & Madani, A. (2023). ProGen2: Exploring the boundaries of protein

language models. *Cell Systems*, *14*(11), 968–978.e3.

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., & et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv Preprint*, arXiv:2311.16452.

Olsen, T. H., Boyles, F., & Deane, C. M. (2022). Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, *31*(1), 141–146.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & et al. (2024). GPT-4 technical report. *arXiv Preprint*, arXiv:2303.08774.

Pan, X., & Shen, H.-B. (2018). Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, *34*(20), 3427–3436.

Patel, A., Singhal, A., Wang, A., Pampari, A., Kasowski, M., & Kundaje, A. (2024). DART-Eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA (version 1). *arXiv Preprint*, arXiv:2412.05430.

Piva, F., Giulietti, M., Burini, A. B., & Principato, G. (2012). SpliceAid 2: A database of human splicing factors expression data and RNA target motifs. *Human Mutation*, *33*(1), 81–85.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., & Ré, C. (2023). Hyena hierarchy: Towards larger convolutional language models. *arXiv Preprint*, arXiv:2302.10866.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2020). Transformer protein language models are unsupervised structure learners. *bioRxiv Preprint*, bioRxiv:2020.12.15.422761.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, *118*(15), e2016239118.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., & et al. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, *13*(1), 1728.

Sato, K., Akiyama, M., & Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, *12*(1), 941.

Schmirler, R., Heinzinger, M., & Rost, B. (2024). Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, *15*(1), 7407.

Shanahan, M. (2024). Talking about LLMs. *Communications of the ACM*, *67*(2), 68–79.

Shen, H., Shen, X., Feng, M., Wu, D., Zhang, C., Yang, Y., Yang, M., Hu, J., Liu, J., Wang, W., & et al. (2022). A universal approach for integrating super large-scale single-cell transcriptomes by exploring gene rankings. *Briefings in Bioinformatics*, *23*(2), bbab573.

Shen, T., Hu, Z., Sun, S., Liu, D., Wong, F., Wang, J., Chen, J., Wang, Y., Hong, L., Xiao, J., & et al. (2024). Accurate RNA 3D structure prediction using a language model-based deep learning approach. *Nature Methods*, *21*(12), 2287–2298.

Simon, E., Swanson, K., & Zou, J. (2024). Language models for biological research: A primer. *Nature Methods*, *21*(8), 1422–1429.

Singh, J., Paliwal, K., Zhang, T., Singh, J., Litfin, T., & Zhou, Y. (2021). Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics*, *37*(17), 2589–2600.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., & et al. (2023). Towards expert-level medical question answering with LLMs. *arXiv Preprint*, arXiv:2305.09617.

Song, T., Dai, H., Wang, S., Wang, G., Zhang, X., Zhang, Y., & Jiao, L. (2022). TransCluster: A cell-type identification method for single-cell RNA-seq data using deep learning based on transformer. *Frontiers in Genetics*, *13*, 1038919.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). ERNIE: Enhanced representation through knowledge integration. *arXiv Preprint*, arXiv:1904.09223.

Tahmid, M. T., Abir, A. R., & Bayzid, M. S. (2024a). BioLLMNet: Enhancing RNA-interaction prediction with a specialized cross-LLM transformation network. *bioRxiv Preprint*, bioRxiv:2024.10.02.616044.

Tahmid, M. T., Shahgir, H. S., Mahbub, S., Dong, Y., & Bayzid, M. S. (2024b). BiRNA-BERT allows efficient RNA language modelling with adaptive tokenization. *bioRxiv Preprint*, bioRxiv:2024.07.02.601703.

Tang, X., Qian, B., Gao, R., Chen, J., Chen, X., & Gerstein, M. (2024). BioCoder: A benchmark for bioinformatics code generation with LLMs. *arXiv Preprint*, arXiv:2308.16458.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. *Nature*, *618*(7965), 616–624.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & et al.. (2023). LLaMA: Open and efficient foundation language model. *arXiv Preprint*, arXiv:2302.13971.

Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., & Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, *7*(1), 39921.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. New York: ACM, 6000–6010.

Wang, N., Bian, J., Li, Y., Li, X., Mumtaz, S., Kong, L., & Xiong, H. (2024a). Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, *6*(5), 548–557.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *arXiv Preprint*, arXiv:2203.11171.

Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., & Gu, Q. (2024b). DPLM-2: A multimodal diffusion protein language model (version 1). *arXiv Preprint*, arXiv:2410.13782.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., & et al. (2022). Emergent abilities of LLMs. *arXiv Preprint*, arXiv:2206.07682.

Wen, H., Tang, W., Dai, X., Ding, J., Jin, W., Xie, Y., & Tang, J. (2023). CellPLM: Pre-training of cell language model beyond single cells. *bioRxiv Preprint*, bioRxiv:2023.10.03.560734.

Wenteler, A., Occhetta, M., Branson, N., Huebner, M., Curean, V., Dee, W. T., Connell, W. T., Hawkins-Hooker, A., Chung, S. P., Ektefaie, Y., & et al. (2024). PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. *bioRxiv Preprint*, bioRxiv:2024.10.02.616248.

Xia, J., Zhao, C., Hu, B., Gao, Z., Tan, C., Liu, Y., Li, S., & Li, S. Z. (2023). Mole-BERT: Rethinking pre-training graph neural networks for molecules. In: *Proceedings of the Eleventh International Conference on Learning Representations*. Appleton: ICLR.

Xu, J., Zhang, A., Liu, F., Chen, L., & Zhang, X. (2023a). CIForm as a transformer-based model for cell-type annotation of large-scale single-cell RNA-seq data. *Briefings in Bioinformatics*, *24*(4), bbad195.

Xu, J., Zhang, A., Liu, F., & Zhang, X. (2023b). STGRNS: An interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics*, *39*(4), btad165.

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., & Yao, J. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, *4*(10), 852−866.

Ye, R., Zhao, H., Wang, X., & Xue, Y. (2024). Technological advancements in deciphering RNA–RNA interactions. *Molecular Cell*, *84*(19), 3722−3736.

Yuan, Y., Chen, Q., & Pan, X. (2024). DGRNA: A long-context RNA foundation model with bidirectional attention Mamba2. *bioRxiv Preprint*, bioRxiv:2024.10.31.621427.

Zhang, Y., Lang, M., Jiang, J., Gao, Z., Xu, F., Litfin, T., Chen, K., Singh, J., Huang, X., Song, G., & et al. (2024). Multiple sequence alignment-based RNA language model and its application to structural inference. *Nucleic Acids Research*, *52*(1), e3.