Large Language Models in Bioinformatics: A Survey

Zhenyu Wang[♡♠*], **Zikang Wang**[♣♡*], **Jiyue Jiang**[♡*], **Pengan Chen**[♢],
Xiangyu Shi[♡], **Yu Li**[♡]
[♡] The Chinese University of Hong Kong, [♠] Peking University Third Hospital,
[♣] The Hong Kong Polytechnic University, [♢] The University of Hong Kong
1810301343@bjmu.edu.cn, zikang.wang@connect.polyu.hk, jiangjy@link.cuhk.edu.hk,
sxysxygm@gmail.com, liyu@cse.cuhk.edu.hk

Abstract

Large Language Models (LLMs) are revolutionizing bioinformatics, enabling advanced analysis of DNA, RNA, proteins, and single-cell data. This survey provides a systematic review of recent advancements, focusing on genomic sequence modeling, RNA structure prediction, protein function inference, and single-cell transcriptomics. Meanwhile, we also discuss several key challenges, including data scarcity, computational complexity, and cross-omics integration, and explore future directions such as multimodal learning, hybrid AI models, and clinical applications. By offering a comprehensive perspective, this paper underscores the transformative potential of LLMs in driving innovations in bioinformatics and precision medicine.

1 Introduction

Bioinformatics is an interdisciplinary field that combines biology, computer science, and information technology to analyze and interpret complex biological data (Abdi et al., 2024). Recently, LLMs have demonstrated remarkable progress in the domain of natural language processing (NLP), with applications that span a wide array of tasks (Min et al., 2023; Raiaan et al., 2024). However, the nature of biological data and the associated tasks differ significantly from text data, presenting unique challenges. The accurate and precise handling of biomedical data to effectively form features and embeddings suitable for LLMs is an ongoing challenge that necessitates innovative solutions.

Within the biological domain, tasks exhibit a high degree of variability and specificity. These include the functional prediction and generation of DNA sequences, the prediction of RNA structure and function, the prediction and design of protein structures, and the analysis of single-cell data, which encompasses dimensionality reduction, clustering, cell annotation, and developmental trajectory analysis. There is a growing interest among researchers in harnessing the power of LLMs for bioinformatics and computational biology, yielding significant results. As illustrated in Figure1, the development, training, and application of large models in bioinformatics are increasing at a rapid pace. Despite this, the diverse methodologies focused on these various tasks have not been systematically summarized and analyzed, presenting an opportunity for comprehensive review and synthesis.

Organization of This Survey: This paper reviews recent advancements in LLMs for bioinformatics. We begin with preliminary concepts (§2), covering key architectures and their relevance to biological data. Next, an overview of representative LLMs in bioinformatics is presented in Table1. We then explore LLM-driven innovations across DNA and Genomics (§3), RNA (§4), Proteins (§5), and Single-Cell Analysis (§6). Finally, we discuss key challenges (§7.1) and propose future directions (§7.2), emphasizing multimodal learning, hybrid AI models, and clinical applications. To conclude, we analyze the limitations of this survey, highlighting areas that require further exploration to fully capture the evolving landscape of LLMs in bioinformatics.

2 Preliminaries

LLMs have demonstrated remarkable advancements across various AI applications, including bioinformatics, where they enable sophisticated sequence modeling, structure prediction, and functional annotation (Li et al., 2024). Mechanistically, the initial design of each LLM typically follows three main architectural paradigms: encoder-only, decoder-only and encoder-decoder models. Each of these architectures has distinct advantages and is suited for different types of bioinformatics tasks.

[*] Equal Contribution

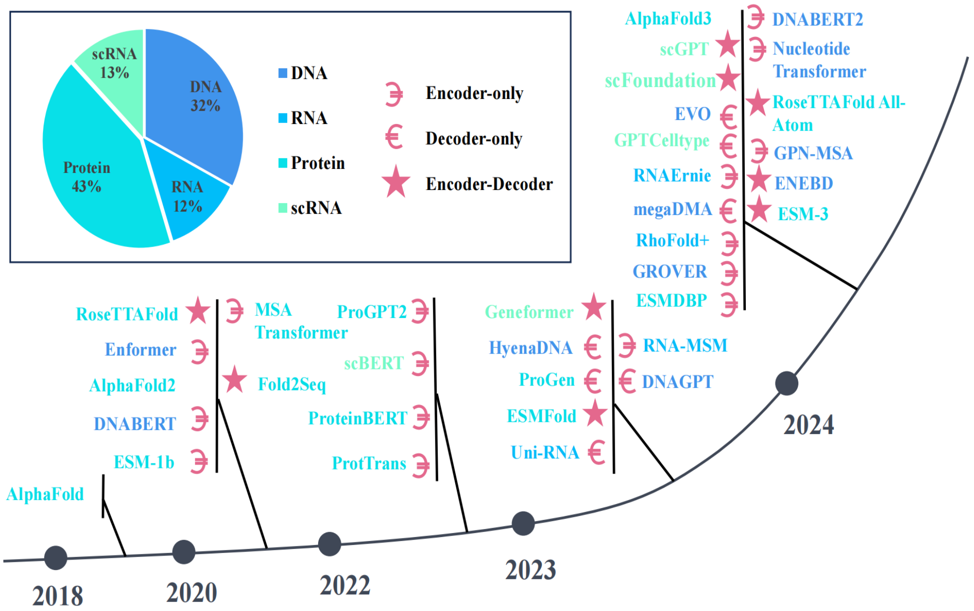

Figure 1: Overview of LLM advancements and applications in bioinformatics, spanning DNA, RNA, protein, and scRNA.

Therefore, in this section, we provide a concise overview of these architectures and corresponding relevance to their applications in bioinformatics.

2.1 encoder-only

Encoder-only models, such as BERT-based architectures (e.g., ProteinBERT (Brandes et al., 2022)), primarily focus on representation learning by capturing contextual dependencies within input sequences. These models utilize bidirectional self-attention, allowing them to learn rich, contextualized embeddings which are crucial for those downstream tasks such as sequence classification, gene expression prediction, and regulatory element identification. However, encoder-only models have limitations in generative tasks, as they lack autoregressive decoding mechanisms.

2.2 decoder-only

Decoder-only models, represented by GPT-based architectures (e.g., ProGen2 (Nijkamp et al., 2023), Evo (Nguyen et al., 2024)), operate in a casual, autogressive manner, which means that they always generate outputs token by token based on previously generated information. These models are particularly well-suited for sequence generation, structure prediction, and functional annotation, making them highly valuable in bioinformatics applications that require de novo sequence synthesis and predictive modeling. Regarding the disadvantages of decoder-only models, for example, their reliance on unidirectional attention can limit their abilities to fully capture long-range bidirectional dependencies, which are essential to understand complicated physiological reactions in vivo. Additionally, they tend to require extensive fine-tuning when applied to domain-specific tasks, as pre-trained general-purpose models may lack sufficient knowledge of biological sequences.

2.3 encoder-decoder

Encoder-decoder models, such as T5-based and transformer-based architectures (e.g., RoseTTAFold (Baek et al., 2021)), are designed for sequence-to-sequence tasks, where an input sequence is transformed into an output sequence. This architecture is particularly useful for the tasks involving mapping between different biological modalities, (e.g., gene expression predication, multiomics data integration). For RoseTTAFold, it employs a three-track neural network to predict protein interactions as well as complex formations. Meanwhile, the encoder-decoder architecture shows great potential when applied to the tasks that require bidirectional context understanding and structured output generation, represented by RNA secondary structure prediction (e.g., RhoFold+ (Shen et al., 2024)) and genome-wide variant effect prediction. However, these models often require substantial computational resources for both training and inference, making them less accessible for researchers with limited computational infrastructure. Additionally, their performance is highly dependent on large-scale domain-specific pre-training, necessitating extensive datasets for generalization.

3 DNA and Genomics: Learn and Generate

The research landscape of LLMs in genomics is witnessing a surge in development, particularly in their application across a spectrum of genomic tasks. These models not only improve the analytical capabilities of DNA sequence analysis, but also accurately predict the impact of genetic mutations, identify key regulatory sequences, and thus advance the understanding of genomic functions. Moreover, the generation of biologically functional gene sequences using LLMs is an area worth exploring, with potential implications for synthetic biology and gene therapy.

The integration of LLMs in genomics has opened new avenues for research, offering insights into the complex interplay between genetic sequences and their biological functions. For instance, in the prediction of gene regulatory elements, LLMs have been shown to outperform traditional machine learning algorithms, providing a more nuanced understanding of gene regulation and expression(Koido et al., 2024). Additionally, their application in the prediction of the effects of genetic mutations has led to a better comprehension of disease mechanisms and potential therapeutic targets.

DNABERT is a pre-trained bidirectional encoder representation, which can capture global and transferrable understanding of genomic DNA sequences based on up and downstream nucleotide contexts. It can be fined tuned to many other sequence analyses tasks such as predict proximal and core promoter regions, identify transcription factor binding sites, recognize canonical and non-canonical splice sites and identify functional genetic variants (Ji et al., 2021).

DNABERT-2 is a high-performance foundation model based on the Transformer architecture, designed for multi-species genome analysis. By incorporating innovative tokenization methods, efficient attention mechanisms, and a comprehensive evaluation benchmark, it provides a powerful tool for understanding and interpreting genomic data across diverse species (Feng et al., 2024).

GeneBERT is a large-scale pre-trained model that integrates diverse genomic data modalities across multiple cell types in a self-supervised manner. By leveraging multi-modal genome data during pre-training, GeneBERT captures complex biological patterns and relationships, enabling robust performance on a wide range of downstream tasks. These tasks include promoter prediction, transcription factor binding site prediction, disease risk estimation, and RNA splicing analysis. Its ability to generalize across various genomic contexts and tasks makes GeneBERT a powerful tool for advancing genome research and precision medicine (Mo et al., 2021).

GROVER is a deep learning model designed to capture the structural and contextual features of DNA sequences by simultaneously learning token-level characteristics and broader sequence contexts. It demonstrates superior performance in both next-k-mer prediction and fine-tuning tasks, such as promoter identification and DNA-protein binding prediction. GROVER's ability to effectively model DNA language structure makes it a valuable tool for advancing genomic research and understanding regulatory mechanisms (Sanabria et al., 2024).

MegaDNA is a groundbreaking long-context generative model designed for genomic sequence analysis and de novo sequence generation. This model leverages a multi-scale Transformer architecture to process and generate DNA sequences at single-nucleotide resolution, enabling unprecedented capabilities in genomics research (Shao and Yan, 2024).

In the realm of gene sequence generation, LLMs hold promise for the creation of sequences with specific biological functions. This capability could facilitate the design of genes for targeted therapies, the enhancement of crop traits, and the development of biomanufacturing processes. The potential of LLMs in this domain is vast, and ongoing research is expected to unveil novel applications and techniques.

Nucleotide Transformer is a state-of-the-art foundation model designed for genomic sequence analysis, leveraging large-scale pre-training on diverse DNA and RNA sequences to capture complex biological patterns. Developed to address the challenges of understanding genomic language, this model integrates advanced deep learning techniques with extensive genomic datasets, enabling robust performance across a wide range of downstream tasks (Dalla-Torre et al., 2024).

Evo is a groundbreaking genomic foundation

Model	Author_Time	Venue	Type	Datasets	Task	Focus
DNABERT	(Ji et al., 2021)	Bioinformatics	Encoder-only	ENCODE	DNA	Predicts genomic functions using DNA sequences
AlphaFold2	Jumper et al., 2021	Nature	-	PDB, BFD, etc	Protein	Predicts protein 3D structures from sequences
RoseTTAFold	Minkyung Baek et al., 2021	Science	Encoder-decoder	PDB, etc	Protein	Predicts protein structures and interactions efficiently
Enformer	Avsec, Ž et al., 2021	Nat Methods	Encoder-only	UCSC Genome Browser, etc	DNA	Predicts gene expression from DNA sequences
ESM-1b	A. Rives, J et al., 2021	PNAS	Encoder-only	UniParc	Protein	Predicts protein structure and function
MSA Transformer	Roshan M Rao et al., 2021	PMLR	Encoder-only	UniRef50	Protein	Predicts protein structures using multiple sequence alignments
Fold2Seq	Yue Cao et al., 2021	PMLR	Encoder-decoder	CATH 4.2	Protein	Designs protein sequences from 3D folds
ProGPT2	Ferruz, N et al., 2022	Nat Commun	Decoder-only	Uniref50	Protein	Generates novel protein sequences with natural-like properties
scBERT	Yang, F et al., 2022	Nat Mach Intell	Encoder-only	Panglao	scRNA	Annotates cell types from single-cell RNA-seq data
ProteinBERT	Nadav Brandes et al., 2022	Bioinformatics	Encoder-only	UniProtKB and UniRef90	Protein	Predicts protein functions from sequences
ProtTrans	A. Elnaggar et al., 2022	IEEE	Encoder-only	UniRef and BFD	Protein	Predicts protein functions and structures
ProGen	Madani, A et al., 2023	Nat Biotechnol	Decoder-only	UniprotKB, etc	Protein	Generates functional protein sequences across families
ESMFold	Zeming Lin et al., 2023	Science	Encoder-decoder	UniRef50	Protein	Predicts protein structures from single sequences
Geneformer	Theodoris, C.V. et al., 2023	Nature	Encoder-decoder	GEO, SRA, HCA, etc	scRNA	Predicts gene networks from single-cell data
HyenaDNA	Eric Nguyen et al., 2023	NeurIPS	Decoder-only	Human reference genome	DNA	Models long-range DNA interactions at single-nucleotide resolution
Uni-RNA	Xi Wang et al., 2023	bioRxiv	Encoder-only	RNAcentral, etc	RNA	Predicts RNA structures, functions, and properties
ProGen2	Erik Nijkamp et al., 2023	Cell Systems	Decoder-only	UniProtKB and BFD	Protein	Generates functional protein sequences across families
xTrimoPGLM	Bo Chen et al., 2023	arXiv	Encoder-decoder	Uniref90 and ColabFoldDB	Protein	Predicts and designs protein sequences and structures
RNA-MSM	Yikun Zhang et al., 2023	Nucleic Acids Research	Encoder-only	RNAcmap	RNA	Predicts RNA structures using evolutionary information
TFBert	Luo H et al., 2023	Interdiscip Sci	Encoder-only	690 ChIP-seq	DNA	Predicts transcription factor binding sites
DNAGPT	Daoan Zhang et al., 2023	arXiv	Decoder-only	GRCh38	DNA	Generates and analyzes DNA sequences
GROVER	Sanabria et al., 2024	Nat Mach Intell	Encoder-only	GRCh37	DNA	Predicts DNA functions from sequence context
megaDNA	Shao, B et al., 2024	Nat Commun	Decoder-only	NCBI GenBank	DNA	Generates and analyzes functional genomes
Nucleotide Transformer	Dalla-Torre, H et al., 2024	Nat Methods	Encoder-only	GRCh38	DNA	Predicts molecular phenotypes from DNA sequences
RNAErnie	Wang, N et al., 2024	Nat Mach Intell	Encoder-only	RNAcentral	RNA	Predicts RNA functions and structures
RhoFold+	Shen, T et al., 2024	Nat Methods	Encoder-only	RNAcentral	RNA	Predicts RNA 3D structures from sequences
GPTCelltype	Hou, W et al., 2024	Nat Methods	Decoder-only	figshare, Zenodo, GEO, etc.	scRNA	Automates cell type annotation using GPT-4
Evo	Eric Nguyen et al., 2024	Science	Decoder-only	OpenGenome	DNA	Predicts and designs DNA, RNA, proteins
scFoundation	Hao, M et al., 2024	Nat Methods	Encoder-decoder	HCA, Single Cell Portal, GEO, etc	scRNA	Predicts and analyzes single-cell transcriptomics data
scGPT	Cui, H et al., 2024	Nat Methods	Encoder-decoder	CELLxGENE	scRNA	Predicts and analyzes single-cell omics data
AlphaFold3	Abramson, J., 2024	Nature	-	PDB	Protein	Predicts biomolecular structures and interactions accurately
DNABERT2	Zhihan Zhou et al., 2024	ICLR	Encoder-only	Human and multi-species genome	DNA	Predicts genomic functions across species efficiently
ESM-DBP	Zeng, W et al., 2024	Nat Commun	Encoder-only	UniProtKB	Protein	Predicts DNA-binding proteins and residues accurately
RoseTTAFold All-Atom	Rohith Krishna et al., 2024	Science	Encoder-decoder	PDB, etc	Protein	Predicts and designs biomolecular structures
ProstT5	Michael Heinzinger et al., 2024	NAR Genomics and Bioinformatics	Encoder-decoder	AFDB	Protein	Translates protein sequences to 3D structures
EpiGePT	Gao, Z et al., 2024	Genome Biol	Encoder-only	ENCODE	DNA	Predicts context-specific epigenomic signals and interactions
RiNALMo	Rafael Josip Penić et al., 2024	arXiv	Encoder-only	RNAcentral	RNA	Predicts RNA structures and functions
ENBED	Aditya Malusare et al., 2024	Bioinformatics Advances	Encoder-decoder	NCBI-Genome	DNA	Analyzes DNA sequences with byte-level precision
GPN-MSA	Benegas, G et al., 2025	Nat Biotechnol	Encoder-only	multiz MSA, etc	DNA	Predicts genome-wide variant effects efficiently
ESM-3	Thomas Hayes et al., 2025	Science	Encoder-decoder	UniRef	Protein	Predicts and designs proteins with multi-modal inputs

Table 1: Overview of representative LLMs in bioinformatics, categorized by architecture, dataset, task, and application domain.

model designed to predict and generate biological sequences—spanning DNA, RNA, and proteins—from molecular to genome scales. Developed by researchers at the Arc Institute and Stanford University, Evo leverages advanced deep learning architectures and large-scale pre-training to achieve unprecedented capabilities in biological sequence analysis and design (Nguyen et al., 2024).

4 RNA: Structure and Function

4.1 RNA Structure

RNA molecules play critical roles in biological systems, functioning as catalytic ribozymes, metabolite-sensing riboswitches, and epigenetically regulatory long noncoding RNAs (Zhang et al., 2022). These diverse functions are enabled by the ability of single-stranded RNA to fold into intricate secondary and tertiary structures. Accurate prediction of RNA structures is therefore essential for understanding their biological mechanisms and therapeutic potential.

However, RNA structure prediction remains challenging due to the complexity and dynamic nature of RNA folding. RNA molecules exhibit conformational flexibility, long-range interactions, and non-canonical base pairing, making their 3D structures difficult to model. Additionally, the scarcity of high-quality experimental RNA structure data limits the training and development of LLMs for this purpose.

Recent advances in computational methods, including deep learning and physics-based simulations, have significantly improved RNA structure prediction. Here, we conclude some RNA-related LLMs recently to better understand developments and limitations in RNA structure predction.

4.1.1 RNA secondary structure prediction

Some researchers benchmarked 6 RNA-LLMs which include RNABERT, RNA-FM, RNA-MSM, ERNIE-RNA, RNAErnie and RiNALMo for their ability to predict RNA secondary structure. They found that RiNALMo and ERNIE-RNA were the models that could better represent and separate the RNA families in the projection without almost overlap. (Zablocki et al., 2024)

4.1.2 RNA Teridry structure prediction

Uni-RNA represents a paradigm shift in RNA research by combining large-scale pre-training with advanced deep learning techniques. Its ability to accurately predict RNA structures, functions, and properties positions it as a powerful tool for accelerating discoveries in RNA biology and therapeutic development (Wang et al., 2023).

RhoFold+ is an advanced method leveraging deep learning and RNA language models to efficiently and accurately predict RNA three-dimensional structures. Its core strength lies in the integration of a large-scale pre-trained RNA language model (RNA-FM) with a deep learning architecture, enabling end-to-end prediction from RNA sequences to 3D structures (Shen et al., 2024).

NuFold is a state-of-the-art deep learning model designed for the accurate prediction of RNA tertiary structures. It addresses the significant gap between RNA sequence data and experimentally determined structures by leveraging advanced computational techniques (Kagaya et al., 2025).

4.2 RNA Function

BEACON comprises 13 distinct tasks derived from extensive prior research, covering structural analysis, functional studies, and engineering applications.Functional tasks focus on the biological roles of RNA, including splice site prediction and noncoding RNA function classification.

RNA-RNA interactions are involved in post-transcriptional processes, contributing to gene expression regulation. RNA-protein interactions are vital for maintaining cellular homeostasis, and disruptions in these interactions can lead to cellular dysfunctions or diseases such as cancer. Furthermore, RNA-small molecule interactions have significant implications in therapeutic development, as RNAs can serve as potential drug targets, especially when conventional protein targets are less accessible.

BioLLMNet provides a way to transform feature space of different molecule's language model features and uses learnable gating mechanism to effectively fuse features and achieves state-of-the-art performance in RNA-protein, RNA-small molecule, and RNA-RNA interactions, outperforming existing methods in RNA-associated interaction prediction (Tahmid et al., 2024).

4.3 RNA sequence generation

RNA-GPT, a multi-modal RNA chat model designed to simplify RNA discovery by leveraging extensive RNA literature. RNA-GPT integrates RNA sequence encoders with linear projection layers and state-of-the-art LLMs for precise representation alignment, enabling it to process user-uploaded

RNA sequences and deliver concise, accurate responses (Xiao et al., 2024).

RNA-DCGen, a generalized framework for RNA sequence generation that is adaptable to any structural or functional properties through straightforward fine-tuning with an RNA language model (RNA-LM).To address these challenges: specialization for fixed constraint types, such as tertiary structures, and lack of flexibility in imposing additional conditions beyond the primary property of interest. (Shahgir et al., 2024)

5 Protein: Prediction and Design

Recently, LLMs have emerged as transformative computational tools in protein research, demonstrating remarkable potential to advance both fundamental comprehension and applied engineering of protein structures and functions. In this section, we classify protein-related LLMs into encoder-only, decoder-only and encoder-decoder architectures based on their diverse protein research applications, such as protein structure and function prediction and protein generation and design.

5.1 Protein Structure and Function

AlphaFold2 employs deep learning to predict protein 3D structures with atomic-level accuracy, achieving unprecedented success in CASP14. Its open-source database has revolutionized structural biology, enabling rapid drug discovery and mechanistic studies across biomedical research(Jumper et al., 2021).

RoseTTAFold integrates sequence, distance, and 3D coordinate predictions through a three-track neural architecture. It also achieves near-experimental accuracy in CASP14, enabling rapid modeling of understudied proteins for therapeutic and evolutionary analyses(Baek et al., 2021).

ESM-1b leverages a transformer-based encoder architecture to infer protein tertiary structures and functional characteristics through self-supervised learning on large-scale protein sequence databases, without relying on manual annotations of the sequence(Meier et al., 2021).

ProteinBERT separates local (character-level) and global (sequence-level) representations and advances transformer architecture through a self-supervised learning paradigm by establishing a unified framework for multitask protein analysis (Brandes et al., 2022).

ProtTrans, a transformer-based protein language model, employs self-supervised learning on 100+ million sequences to capture evolutionary-structural patterns. It achieves state-of-the-art performance in tertiary structure prediction, functional annotation, and engineering design while enabling efficient transfer learning across diverse proteomic tasks(Elnaggar et al., 2021).

AlphaFold3 advances structural biology by integrating geometric deep learning with diffusion models, achieving atomic-resolution predictions for generalized biomolecular complexes (proteins, DNA, ligands). It demonstrates better accuracy in ligand binding sites over experimental methods, revolutionizing drug discovery and systems biology through whole-cell interactome modeling (Abramson et al., 2024).

ESM-DBP integrates protein language models with DNA-binding specificity prediction, leveraging evolutionary-scale sequence training to accurately identify DNA-interaction motifs and binding sites(Zeng et al., 2024).

RoseTTAFold All-Atom is a fast and accurate neural network for predicting diverse biomolecular assemblies. It models proteins, nucleic acids, small molecules, metals, and covalent modifications, advancing structural biology and drug discovery (Krishna et al., 2024).

5.2 Protein Design and Engineering

LLMs will be well suited for protein design applications, such as designing antibodies with reduced aggregation propensity, development of drugs that target specific protein phases in diseases, and understanding the mechanisms behind diseases caused by protein misfolding and aggregation.

ProtGPT2 is a pretrained transformer-based language model for protein sequence generation and engineering. It explores new protein regions while preserving natural protein features, offering high-throughput design capabilities (Ferruz et al., 2022).

ProGen2 is a protein language models with 6.4B parameters, trained on over a billion proteins. It excels in capturing sequence distribution, generating novel sequences, and predicting protein fitness without fine-tuning (Nijkamp et al., 2023).

ESM-3 is a large language model trained on protein sequences, structures, and functions. It offers multimodal analysis, generating novel proteins and predicting 3D structures, advancing drug discovery and biotechnology (Hayes et al., 2025).

6 scRNA: Development and Challenge

Single-cell sequencing technology, an evolution of transcriptome sequencing, enables the examination of gene expression and transcription levels at the individual cell level (Potter, 2018). This technology is pivotal for understanding various biological processes at the cellular level, such as disease progression, therapeutic efficacy, and resistance. It can identify relevant cell subpopulations and molecules, holding profound significance for the advancement of bioinformatics. However, the effective and accurate processing and analysis of vast single-cell data sets present a challenge for bioinformaticians. Traditional analysis methods and pipelines are based on tools like Seurat and Scanpy (Hao et al., 2024b) (Wolf et al., 2018). With the advancement of deep learning, numerous studies have integrated frameworks such as deep learning and large language models with the processing and analysis of single-cell data, achieving significant progress and demonstrating the immense potential of combining these approaches to faster development in the field of single-cell research (Molho et al., 2024).

scBERT is a pioneering deep learning model designed for cell type annotation in scRNA-seq data. It adapts the BERT (Bidirectional Encoder Representations from Transformers) framework, originally developed for NLP, to the domain of single-cell transcriptomics. The model leverages a pre-training and fine-tuning paradigm, where it first learns general patterns of gene-gene interactions from large-scale unlabeled scRNA-seq data and then fine-tunes on specific tasks using labeled data to predict cell types (Molho et al., 2024).

Geneformer is a foundational deep learning model based on the Transformer architecture, specifically designed for analyzing scRNA-seq data. Pre-trained in a massive corpus of 29.9 million single-cell human transcriptomes, Geneformer uses self-supervised learning to capture gene-gene interactions and regulatory dynamics in various cell types and tissues (Theodoris et al., 2023).

GPTCelltype is an innovative R software package designed to automate cell type annotation in scRNA-seq analysis by leveraging the capabilities of GPT-4, a state-of-the-art LLMs. This tool represents a significant advancement in the field of bioinformatics, offering a cost-effective and efficient alternative to traditional manual or semi-automated annotation methods (Hou and Ji, 2024).

scFoundation is a groundbreaking large-scale pre-trained model designed for scRNA-seq data analysis, featuring 100 million parameters and capable of handling approximately 20,000 genes simultaneously. Pre-trained in more than 50 million human single-cell transcriptomic profiles, scFoundation represents a significant advancement in the field of single-cell genomics, offering a robust framework for diverse downstream tasks such as gene expression enhancement, drug response prediction, and cell type annotation (Hao et al., 2024a).

scGPT is a state-of-the-art foundation model designed for single-cell multi-omics data analysis, leveraging the transformer architecture to capture complex gene-cell interactions. Pre-trained in over 33 million human single-cell transcriptomes from the CELLxGENE database, scGPT excels in extracting meaningful biological insights and generalizing across diverse downstream tasks, such as cell type annotation, perturbation prediction, batch integration, and gene regulatory network inference (Cui et al., 2024).

Foundation models like scBERT, Geneformer, and scGPT have transformed single-cell data analysis through self-supervised pretraining on millions of transcriptomes, enabling robust transfer learning for tasks such as cell annotation, perturbation prediction, and batch correction. Architectural innovations (e.g., performer encoders, rank-based encoding) address computational challenges of high-dimensional data, while task-specific designs improve generalizability and interpretability via attention mechanisms.

However, limitations persist, including trade-offs between computational efficiency and expression resolution, biases from imbalanced pretraining data, and the "black-box" nature of Transformer architectures. High computational costs further hinder accessibility. Future efforts should prioritize multimodal integration of transcriptomic, epigenetic, and spatial data, knowledge-guided architectures incorporating biological networks, efficient few-shot learning frameworks, and lightweight adaptations for greater usability. Addressing these challenges will enhance biological interpretability and bridge AI capabilities with actionable insights, advancing precision medicine and systems biology.

7 Conclusions and Future Directions

In conclusion, this paper comprehensively examined the applications of LLMs across DNA, RNA, protein, and single-cell data analysis, highlighting key research contributions and emerging methodologies. Despite advances, LLM applications in bioinformatics remain evolving, requiring key challenges to address for full potential. Therefore, we here discuss the current limitations and outline promising future directions for advancing LLM-driven bioinformatics research.

7.1 Challenges and Limitations

7.1.1 Data Scarcity, Quality and Bias

LLMs always require large-scale, high-quality biological datasets for effective training, yet annotated genomic, transcriptomic, and proteomic data remain limited (Lu et al., 2024). Unlike natural language corpora, which are abundant and diverse, biological datasets are often noisy, incomplete, or biased toward well-studied species and diseases. Consequently, model generalizability suffers, leading to biased predictions which may not hold across diverse biological contexts. Additionally, batch effects and experimental noise complicate the development of robust foundation models for bioinformatics (Yu et al., 2024).

7.1.2 Computational Complexity and Model Efficiency

State-of-the-art LLMs, such as AlphaFold and DNABERT, require massive computational resources for both training and inference. This computational barrier limits accessibility, particularly for research groups with limited infrastructure. Furthermore, relatively longer biological sequences significantly increase memory and processing requirements, making it challenging to apply standard Transformer architectures to genome-scale data (Bernard et al., 2025). Efficient methods of model compression and retrieval-augmented need to be further explored to enhance scalability.

7.1.3 Multimodal Learning and Cross-Omics Integration

Biological systems exhibit intricate interactions across multiple molecular layers, including but not limited to genomics and metabolomics. Despite recent advancements, current LLMs remain predominantly trained on single-modality datasets, constraining their ability to model cross-scale molecular dependencies. Addressing this limitation requires the development of multimodal architectures capable of integrating heterogeneous biological data in a biologically meaningful and computationally efficient manner (Dankan Gowda et al., 2025).

7.2 Future Directions

To overcome the challenges mentioned above, future research should focus on developing efficient, interpretable, and multimodal LLM architectures tailored for bioinformatics.

7.2.1 Hybrid AI Models for Biological Reasoning

Integrating LLMs with mechanistic models, represented by graph neural networks (GNNs), and knowledge graphs could improve biological reasoning and interpretability(Feng et al., 2025). Hybrid approaches that combine deep learning with symbolic AI(Colelough and Regli, 2025) or constraint-based modeling(Bystrova et al., 2024) may enable causality-aware predictions in biological systems.

7.2.2 Multimodal and Cross-Omics Integration

Future LLMs should be designed with multimodal learning capabilities, enabling the simultaneous processing of DNA, RNA, protein, and epigenetic data. By integrating self-supervised learning across diverse omics datasets, these models could enhance biological understanding and improve cross-species generalization. Furthermore, to enhance explainability, biologically constrained LLMs incorporating evolutionary principles and regulatory networks could improve model robustness and reliability(Feng et al., 2023), ensuring greater alignment with fundamental biological mechanisms.

7.2.3 Towards Clinical and Biomedical Applications

Bridging the gap between AI-driven bioinformatics and real-world biomedical applications necessitates further advancements in model validation, regulatory compliance, and ethical considerations. Moving forward, LLM tools require rigorous clinical evaluation and experimental benchmarking to ensure healthcare reliability and safety.(Perlis and Fihn, 2023).

In summary, addressing these challenges and research directions will enable next-generation LLMs to drive transformative breakthroughs in genomics and precision medicine, paving the way for a new era of AI-driven biological discovery.

Limitations

In this paper, we provide a survey of LLMs in bioinformatics. Despite our best efforts, there may be still several limitations that remain in this study.

Scope Restriction: Our survey focuses on four subdomains: DNA, RNA, protein, and single-cell analysis. However, other areas, represented by epigenomics and metagenomics, are not covered in depth. Expanding the scope to include these fields would provide a more holistic perspective on LLM-driven bioinformatics research.

Rapidly Evolving Field: Given the rapid advancements in LLMs, some recent breakthroughs may not be fully captured in this survey. The field is continuously evolving, with new models and methodologies emerging at a fast pace.

Lack of Empirical Benchmarking: While this paper synthesizes and analyzes existing research, it does not include direct experimental validation or standardized benchmarking of LLMs for bioinformatics tasks. Specifically, a rigorous assessment of LLMs' performance under consistent conditions, such as dataset standardization and computational efficiency, remains an open challenge.

Moving forward, we will actively monitor and incorporate emerging developments to ensure that future iterations of this work reflect the most up-to-date progress in LLMs applications for bioinformatics.

Ethical Statement

There are no ethical issues.

References

Ghlomareza Abdi, Mukul Jain, Mukul Barwant, Reshma Tendulkar, Mugdha Tendulkar, Mohd Tariq, and Asad Amir. 2024. Unveiling the dynamic role of bioinformatics in automation for efficient and accurate data processing and interpretation. In *Advances in Bioinformatics*, pages 279–319. Springer.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876.

Clément Bernard, Guillaume Postic, Sahar Ghannay, and Fariza Tahi. 2025. Rna-torsionbert: leveraging language models for rna 3d torsion angles prediction. *Bioinformatics*, 41(1):btaf004.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Daria Bystrova, Charles Assaad, Julyan Arbel, Emilie Devijver, Éric Gaussier, and Wilfried Thuiller. 2024. Causal discovery from time series with hybrids of constraint-based and noise-based algorithms. *Transactions on Machine Learning Research Journal*.

Brandon C Colelough and William Regli. 2025. Neurosymbolic ai in 2024: A systematic review. *arXiv preprint arXiv:2501.05435*.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. 2024. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, pages 1–11.

V Dankan Gowda, D Palanikkumar, KDV Prasad, Mandeep Kaur, and Shivoham Singh. 2025. Future directions and emerging trends in multimodal data fusion for bioinformatics. *Multimodal Data Fusion for Bioinformatics Artificial Intelligence*, pages 247–282.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127.

Ke Feng, Hongyang Jiang, Chaoyi Yin, and Huiyan Sun. 2023. Gene regulatory network inference based on causal discovery integrating with graph neural network. *Quantitative Biology*, 11(4):434–450.

Tao Feng, Yi Huang, and Beiyu Li. 2024. LLM-based DNA promoter prediction. In *Submitted to CS582 ML for bioinformatics workshop*. Under review.

Yichun Feng, Lu Zhou, Chao Ma, Yikai Zheng, Ruikun He, and Yixue Li. 2025. Knowledge graph–based thought: a knowledge graph–enhanced llm framework for pan-cancer question answering. *GigaScience*, 14:giae082.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024a. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pages 1–11.

Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. 2024b. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2):293–304.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. 2025. Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pages 1–4.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Yuki Kagaya, Zicong Zhang, Nabil Ibtehaz, Xiao Wang, Tsukasa Nakamura, Pranav Deep Punuru, and Daisuke Kihara. 2025. Nufold: end-to-end approach for rna tertiary structure prediction with flexible nucleobase center representation. *Nature Communications*, 16(1):881.

Masaru Koido, Kohei Tomizuka, and Chikashi Terao. 2024. Fundamentals for predicting transcriptional regulations from dna sequence patterns. *Journal of Human Genetics*, pages 1–6.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. 2024. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528.

Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Gengjie Jia, Sheng Wang, Le Song, and Yu Li. 2024. Progress and opportunities of foundation models in bioinformatics. *Briefings in Bioinformatics*, 25(6):bbae548.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024. Large language models in biomedicine and health: current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9):1801–1811.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. 2021. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Shentong Mo, Xi Fu, Chenyang Hong, Yizhen Chen, Yuxuan Zheng, Xiangru Tang, Yanyan Lan, Zhiqiang Shen, and Eric Xing. 2021. Multi-modal self-supervised pre-training for large-scale genome data. In *NeurIPS 2021 AI for Science Workshop*.

Dylan Molho, Jiayuan Ding, Wenzhuo Tang, Zhaoheng Li, Hongzhi Wen, Yixin Wang, Julian Venegas, Wei Jin, Renming Liu, Runze Su, et al. 2024. Deep learning in single-cell analysis. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–62.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. 2023. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978.

Roy H Perlis and Stephan D Fihn. 2023. Evaluating the application of large language models in clinical research contexts. *JAMA Network Open*, 6(10):e2335924–e2335924.

S Steven Potter. 2018. Single-cell rna sequencing for the study of development, physiology and disease. *Nature Reviews Nephrology*, 14(8):479–492.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Melissa Sanabria, Jonas Hirsch, Pierre M Joubert, and Anna R Poetsch. 2024. Dna language model grover learns sequence context in the human genome. *Nature Machine Intelligence*, 6(8):911–923.

Haz Sameen Shahgir, Md Rownok Zahan Ratul, Md Toki Tahmid, Khondker Salman Sayeed, and Atif Rahman. 2024. Rna-dcgen: Dual constrained rna sequence generation with llm-attack. *bioRxiv*, pages 2024–09.

Bin Shao and Jiawei Yan. 2024. A long-context language model for deciphering and generating bacteriophage genomes. *Nature Communications*, 15(1):9392.

Tao Shen, Zhihang Hu, Siqi Sun, Di Liu, Felix Wong, Jiuming Wang, Jiayang Chen, Yixuan Wang, Liang Hong, Jin Xiao, et al. 2024. Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods*, pages 1–12.

Md Toki Tahmid, Abrar Rahman Abir, and Md Shamsuzzoha Bayzid. 2024. Biollmnet: Enhancing rna-interaction prediction with a specialized cross-llm transformation network. *bioRxiv*, pages 2024–10.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. 2023. Uni-rna: universal pre-trained models revolutionize rna research. *bioRxiv*, pages 2023–07.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. 2018. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5.

Yijia Xiao, Edward Sun, Yiqiao Jin, and Wei Wang. 2024. Rna-gpt: Multimodal generative system for rna sequence understanding. *arXiv preprint arXiv:2411.08900*.

Ying Yu, Yuanbang Mai, Yuanting Zheng, and Leming Shi. 2024. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology*, 25(1):254.

LI Zablocki, LA Bugnon, M Gerard, L Di Persia, G Stegmayer, and DH Milone. 2024. Comprehensive benchmarking of large language models for rna secondary structure prediction. *arXiv preprint arXiv:2410.16212*.

Wenwu Zeng, Yutao Dou, Liangrui Pan, Liwen Xu, and Shaoliang Peng. 2024. Improving prediction performance of general protein language model by domain-adaptive pretraining on dna-binding protein. *Nature Communications*, 15(1):7838.

Jinsong Zhang, Yuhan Fei, Lei Sun, and Qiangfeng Cliff Zhang. 2022. Advances and opportunities in rna structure experimental determination and computational modeling. *Nature methods*, 19(10):1193–1207.