Large Language Models in Bioinformatics: A Survey

www.surveyx.cn

Abstract

Large Language Models (LLMs) have emerged as pivotal tools in bioinformatics, offering transformative advancements in processing extensive biological data. This survey paper explores the multifaceted applications of LLMs, particularly in sequence analysis and protein structure prediction, which are crucial for genomics and proteomics. LLMs enhance model evaluations and address challenges such as data leakage, offering unified search methods in rapidly growing protein databases. Despite their potential, challenges persist, including data quality, computational demands, and model interpretability. The integration of LLMs in bioinformatics necessitates domain-specific adaptations to address these challenges effectively. This survey systematically evaluates LLMs' performance, highlighting their transformative potential and limitations. The scope includes historical developments, applications in clinical settings, and ethical considerations. By focusing on the limitations and ethical challenges, the survey provides a balanced perspective crucial for informing future research. It concludes by emphasizing the need for robust evaluation frameworks and ethical guidelines to harness LLMs' capabilities responsibly, ultimately enhancing bioinformatics research and applications.

1 Introduction

1.1 Concept of Large Language Models (LLMs)

Large Language Models (LLMs) are advanced AI systems that significantly enhance natural language processing (NLP) capabilities across various applications, including chatbots and virtual assistants [1]. Characterized by their extensive parameter sizes, LLMs effectively understand and generate human language [2]. These models have become integral to artificial intelligence, showcasing robust abilities in both natural language understanding and generation [3].

The evolution of LLMs, transitioning from statistical models to sophisticated transformer-based architectures, reflects major advancements in design and training methodologies [4]. This progression positions LLMs as crucial components in the AI landscape, categorizing methodologies and addressing their roles across various domains [5].

LLMs are particularly valuable in constructing complex software models, especially in bioinformatics, where they analyze large datasets and facilitate intricate computational tasks [6]. They represent a significant advancement in AI, moving beyond traditional sequence modeling methods like RNNs and Transformers to efficiently manage long sequences [7]. Trained on extensive datasets, LLMs demonstrate versatility and transformative potential in language understanding and generation [8].

1.2 Relevance of LLMs in Bioinformatics

LLMs are increasingly vital in bioinformatics, driving transformative advancements by improving model evaluations and addressing challenges such as data leakage in performance assessments [9]. They excel in processing vast biological data, particularly in genomics, proteomics, and drug discovery, enhancing the understanding of protein sequences and their related textual information, which is essential for advancing bioinformatics research [10]. The rapid growth of protein-related

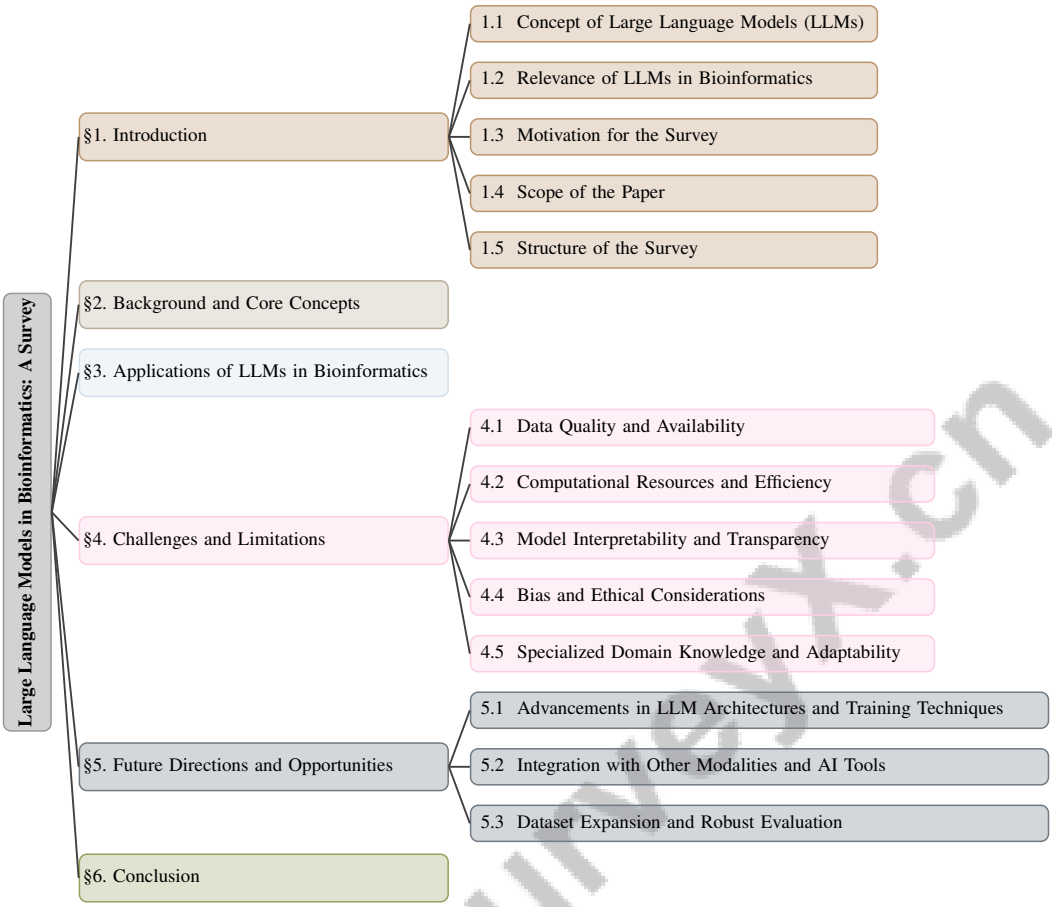


Figure 1: chapter structure

databases and scientific literature necessitates unified search methods, effectively addressed by LLMs [11].

In proteomics, LLMs automate complex research tasks, reducing human bias and improving data analysis efficiency [12]. They utilize evolutionary information from protein sequences, providing efficient management of the expanding protein sequence databases [13]. Additionally, LLMs help bridge the sequence-structure gap in computational biology, enabling innovative methods for predicting protein structures and functions from vast amounts of unlabeled sequence data [14]. The RLCS problem has significant implications in bioinformatics, particularly for identifying similarities and discovering patterns among DNA, RNA, and protein sequences [15].

Despite their potential, integrating LLMs into bioinformatics faces challenges. While proficient in processing general knowledge, LLMs often lack the precision needed for generating domain-specific insights [16]. This limitation highlights the need for continuous refinement of LLMs for specialized bioinformatics contexts. The survey emphasizes that Transformer-based language models can bridge knowledge gaps between NLP and bioinformatics, enhancing research capabilities [17]. Although the use of LLMs in bioinformatics is still developing, they hold promise for addressing historical challenges like data scarcity and noise.

The parametric domain knowledge embedded within LLMs is accelerating their adoption in real-world biomedical applications. For instance, LLMs can generate insightful definitions for biomedical concepts, enriching training data for semantic models [18]. Their application in healthcare settings enhances the accuracy and efficiency of medical tasks, improving patient engagement and educational outcomes [19]. However, challenges persist, as models like GPT-4 struggle with domain-specific issues, leading to poor generalization and inaccuracies [20]. Despite these challenges, LLMs present substantial opportunities for advancing bioinformatics research, necessitating careful consideration of their limitations, including potential for nonsensical outputs and misinterpretation of intent [16].

1.3 Motivation for the Survey

This survey is motivated by the need to systematically explore the potential and limitations of LLMs in bioinformatics, a field rapidly transforming due to advancements in artificial intelligence. As LLMs evolve, they present both opportunities and challenges in clinical practice, medical research, and education, necessitating a thorough assessment of their impact and implications [21]. The survey aims to bridge existing knowledge gaps by providing a comprehensive overview of advancements, applications, and ethical considerations associated with LLMs in bioinformatics [8].

The integration of LLMs in bioinformatics offers promising solutions to the complexities of high-throughput omics data and traditional hypothesis-driven methodologies, particularly in automating bioinformatics workflows, which include tasks like literature searching, code execution, and data analysis [22]. By systematically evaluating LLM performance, this survey seeks to facilitate comparisons among different models and advance AI applications within the life sciences [21].

Furthermore, the survey addresses the necessity for specialized evaluation frameworks to ensure the ethical and effective deployment of LLMs in medical and biological research [4]. By highlighting both the transformative potential and limitations of LLMs, the survey provides a balanced perspective critical for informing future research and applications in bioinformatics. Through this comprehensive review, the survey aims to illuminate key milestones in the integration of AI/ML technologies in biological research, incorporating both technical and ethical dimensions [23].

1.4 Scope of the Paper

This survey defines the boundaries of LLM applications in bioinformatics, emphasizing their transformative role across various omics levels, including genomics, proteomics, and transcriptomics [16]. It encompasses the historical development of computing in life sciences, tracing the evolution from early computational models to contemporary AI applications [24]. The survey examines LLMs in clinical settings, medical text data processing, and public health awareness, while intentionally excluding non-medical applications and detailed technical aspects of specific model architectures [23].

The survey systematically addresses various types of LLMs, including task-based financial LLMs, multilingual LLMs, biomedical and clinical LLMs, vision language models, and code language models, providing a broad overview of their applications [4]. Additionally, it discusses the integration of AI concepts and methodologies, focusing on Networks, Layers, Functions, LLMs, Preprocessing, and Bias [5]. Major aspects of LLMs, including architectures, pre-training objectives, transfer learning methods, and scalability challenges, are also covered [8].

In bioinformatics, the survey explores the application of Transformer models for nucleotide sequences, including promoter prediction, methylation analysis, read classification, and binding predictions [17]. It examines the modular architecture of BRAD, which integrates various bioinformatics tools and LLMs, enabling a more flexible workflow [22]. The survey excludes discussions on unrelated AI applications outside the biological context, ensuring a focused examination of the challenges and ethical considerations surrounding LLMs [25].

Focusing on the limitations, harms, and risks associated with LLMs, this survey excludes discussions on other AI technologies not directly related to LLMs, ensuring a concentrated examination of the challenges and ethical considerations inherent in their use. Additionally, while foundational State Space Models (SSMs) are included, detailed discussions on non-SSM methods like classical RNNs and LSTMs are excluded [6].

1.5 Structure of the Survey

This survey is systematically organized into several key sections, each focusing on distinct aspects of LLMs and their applications in bioinformatics. The structure is inspired by the evolution of language models as outlined by Zhao et al. [2], divided into four major stages: statistical language models, neural language models, pre-trained language models, and large language models. Each stage emphasizes scaling laws and emergent abilities, offering a comprehensive framework for understanding LLM progression and capabilities.

The initial section introduces LLMs, highlighting their relevance and transformative potential in bioinformatics. It establishes the foundation by defining LLMs, discussing their role, and outlining the motivation and scope of the survey. This is followed by an exploration of the background and core concepts, elaborating on the intersection of bioinformatics, computational biology, and artificial intelligence. Key concepts such as sequence analysis, protein structure prediction, and genomics are defined, emphasizing how LLMs enhance AI applications in biology.

Subsequent sections delve into specific applications of LLMs in bioinformatics, focusing on drug discovery, biomedical research, clinical settings, and personalized medicine. Each application is examined through examples and case studies, illustrating the practical utility of LLMs in advancing bioinformatics research.

The survey provides a detailed examination of the challenges and limitations associated with LLM applications in bioinformatics. Key issues include concerns regarding data quality, which can impact model output accuracy; the significant computational resources required for training and deploying these models; complexities surrounding model interpretability that hinder decision-making understanding; and ethical considerations related to bias and potential misuse of generated content. This comprehensive overview aims to inform researchers and practitioners about critical factors necessary to enhance LLM effectiveness and reliability in bioinformatics applications [21, 26, 27, 23, 4]. These challenges are critically analyzed to provide a balanced perspective on LLM deployment in the field.

Finally, the survey concludes with a discussion on future directions and opportunities for LLMs in bioinformatics. The ongoing evolution and potential of LLMs in bioinformatics research are underscored by emerging trends and technological advancements, revealing their transformative capabilities across various omics levels. This highlights not only LLM applications in enhancing data analysis and scientific communication but also the challenges they pose regarding reproducibility and authenticity in scientific publications. Insights into these developments suggest that LLMs could significantly improve bioinformatics research efficiency, enabling researchers with limited programming expertise to leverage advanced coding tools and contribute more effectively to the field [27, 23, 8].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Intersection of Bioinformatics, Computational Biology, and AI

The intersection of bioinformatics, computational biology, and artificial intelligence (AI) enhances the analysis and interpretation of complex biological data. This synergy is evident in sequence analysis, where the Restricted Longest Common Subsequence (RLCS) problem extends the Longest Common Subsequence (LCS) problem, crucial for identifying similarities in DNA, RNA, and protein sequences [15]. Such analyses elucidate biological functions and evolutionary relationships. Integrating AI into bioinformatics automates data processing, addressing challenges from the rapid expansion of protein databases and scientific literature [11]. AI-driven tools, like Large Language Models (LLMs), enable efficient synthesis and interpretation of data, driving innovations in genomics, proteomics, and personalized medicine. This convergence not only accelerates discovery but also fosters novel computational methods to tackle complex biological questions, advancing life sciences and improving human health.

2.2 Key Concepts: Sequence Analysis, Protein Structure Prediction, and Genomics

Core bioinformatics concepts—sequence analysis, protein structure prediction, and genomics—are significantly enhanced by LLMs [27]. Sequence analysis examines DNA, RNA, or protein sequences to elucidate structure, function, and evolution, essential for identifying genetic variations and understanding biological processes. A critical challenge is predicting the likelihood of unseen nucleotide sequences belonging to human genes and generating new sequences for candidate genes [28]. Protein structure prediction focuses on determining three-dimensional shapes from amino acid sequences, a complex task due to intricate folding patterns influenced by sequences and evolutionary information [14]. Accurate predictions are vital for understanding protein functions and interactions. Genomics, the comprehensive study of an organism's DNA, is crucial for exploring the genetic basis of diseases and traits, involving predictions of gene properties, including regulatory elements and protein func-

tions [29]. The field is further complicated by the diverse functions and regulatory mechanisms of non-coding RNAs like lncRNAs, largely unexplored [30]. LLMs address several biological problems, including domain exploration, sequence analysis, structure construction, function prediction, and multimodal integration [25]. Challenges persist, such as the computational expense of Transformer models and the need for tailored models for specific contexts [17]. Despite hurdles, LLMs offer promising advancements in bioinformatics by facilitating accurate and efficient analyses of complex biological data.

2.3 Enhancing AI Applications in Biology with LLMs

LLMs have advanced AI applications in biological research by providing scalable models excelling in language understanding and generation, thus facilitating complex biological data analysis [8]. Their integration into bioinformatics workflows, as demonstrated by systems like BRAD, enhances research efficiency and effectiveness [22]. LLMs improve molecular analysis precision and deepen understanding of genetic-phenotype interactions through domain-specific models that merge insights from LLMs with biomedical knowledge graphs. BioLORD-2023, for instance, advances semantic textual similarity and biomedical concept representation [18]. Frameworks like TourSynbio-Search, a protein multimodal large language model, illustrate how LLMs facilitate unified searches in protein engineering, advancing AI applications [11]. In protein sequence analysis, models like PEvoLM enhance analytical capabilities by predicting the next amino acid in a sequence and the corresponding PSSM column distribution [13]. LLMs also play a vital role in genetic-phenotype knowledge representation, integrating genomic sources to elucidate genetic relationships [31]. Challenges remain in optimizing LLMs for biological applications, particularly in managing long sequences due to the quadratic attention complexity and memory requirements of traditional models like Transformers [7]. Addressing these involves constructing robust datasets to enhance LLM performance, ensuring their continued contribution to advancing AI applications in biology [3]. Effective prompt engineering and user interaction are crucial for realizing LLMs' full potential in research [32]. Ongoing efforts, such as Self-BioRAG, aim to refine LLMs for specific contexts by incorporating specialized retrieval techniques and self-reflection mechanisms, improving their biomedical performance [20]. The versatility and adaptability of LLMs across bioinformatics tasks are underscored by frameworks categorizing existing research based on application contexts, highlighting their transformative potential [17].

In recent years, the application of Large Language Models (LLMs) in bioinformatics has garnered significant attention due to their transformative potential across various domains. As illustrated in Figure 2, the hierarchical classification of LLM applications in this field encompasses critical areas such as drug discovery, biomedical research, clinical practice, and personalized medicine. This figure not only delineates these categories but also highlights essential components including data analysis, proteomics, genomics, and patient engagement. By emphasizing the role of LLMs in enhancing efficiency, accuracy, and personalization within healthcare, the figure underscores the profound impact these models can have on advancing bioinformatics and improving patient outcomes.

3 Applications of LLMs in Bioinformatics

3.1 LLMs in Drug Discovery and Biomedical Research

Large Language Models (LLMs) have become instrumental in drug discovery and biomedical research by enhancing data analysis and interpretation. Their integration improves efficiency and accuracy, facilitating innovative therapeutic strategies. For instance, PEvoLM efficiently encodes evolutionary information, aiding in drug target identification and understanding disease mechanisms [13]. In proteomics, tools like PROTEUS automate analysis pipelines, reducing human bias and expediting the assessment of protein interactions, crucial for discovering new drug candidates [12]. Similarly, GP-GPT excels in genomics tasks, elucidating genetic interactions and disease pathways, thereby advancing personalized medicine [31].

The categorization of LLM applications in drug discovery and biomedical research is illustrated in Figure 3, which highlights key tools and systems such as PEvoLM, PROTEUS, and GP-GPT in drug discovery, alongside ProtTrans, OncoNet Ontology, and TourSynbio-Search in biomedical research, as well as Lomics, Self-BioRAG, and BRAD in information retrieval. This visual representation underscores the diverse functionalities of LLMs across various domains.

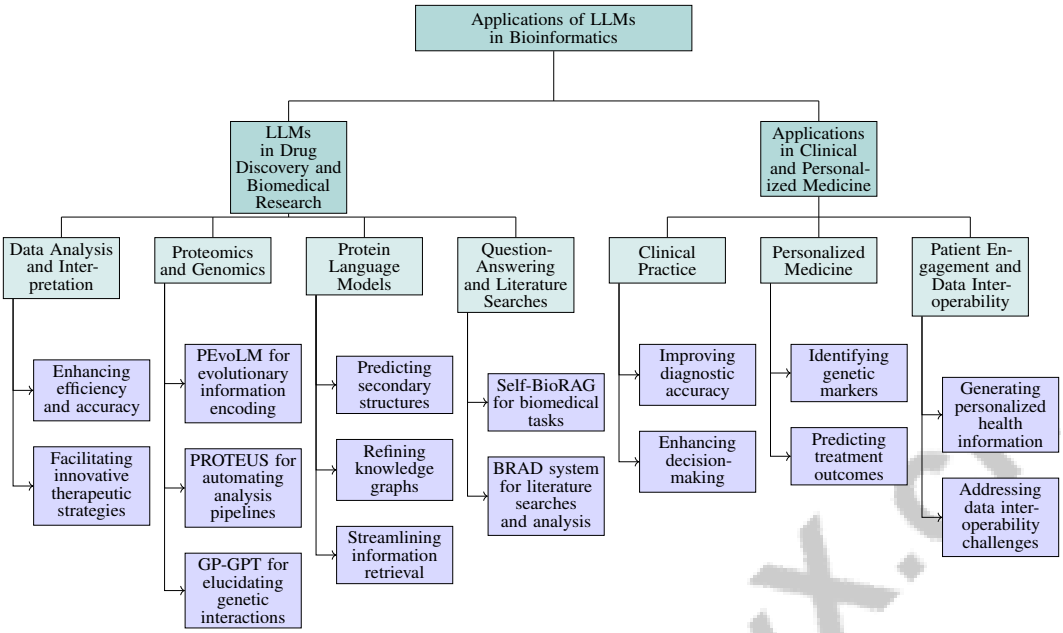


Figure 2: This figure illustrates the hierarchical classification of Large Language Models (LLMs) applications in bioinformatics, focusing on drug discovery, biomedical research, clinical practice, and personalized medicine. It highlights key areas such as data analysis, proteomics, genomics, and patient engagement, emphasizing LLMs' role in improving efficiency, accuracy, and personalization in healthcare.

Protein language models contribute to drug discovery by predicting secondary structures and subcellular localization, essential for understanding protein behavior during drug design [14]. LLMs also refine knowledge graphs with recent scientific literature, enhancing research accuracy [33]. Information retrieval is streamlined by tools like TourSynbio-Search, bridging complex biological databases and researchers [11]. Lomics enhances pathway analysis specificity, accelerating the identification of relevant pathways for drug targets [34].

Moreover, LLMs improve question-answering capabilities, as demonstrated by Self-BioRAG, which outperforms state-of-the-art models in biomedical tasks [20]. The BRAD system exemplifies LLM utility in literature searches, gene enrichment analysis, and software execution, enhancing research efficiency in drug discovery and biomedical research [22].

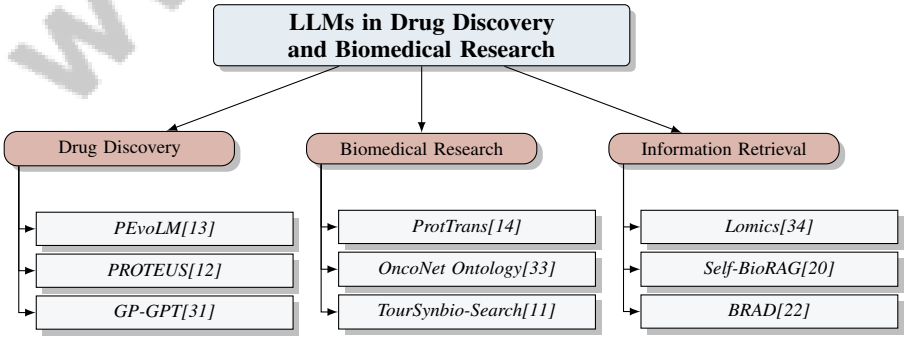
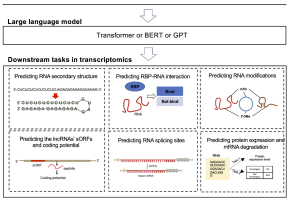


Figure 3: This figure illustrates the categorization of Large Language Models (LLMs) applications in drug discovery and biomedical research, highlighting key tools and systems like PEvoLM, PROTEUS, and GP-GPT in drug discovery, ProtTrans, OncoNet Ontology, and TourSynbio-Search in biomedical research, and Lomics, Self-BioRAG, and BRAD in information retrieval.

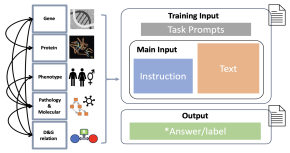
3.2 Applications in Clinical and Personalized Medicine

LLMs are increasingly integrated into clinical settings and personalized medicine, enhancing precision and efficiency in healthcare delivery. They analyze complex medical data to develop tailored treatment strategies for individual patients. In clinical practice, LLMs improve diagnostic accuracy and decision-making by processing vast medical literature and patient records, offering evidence-based insights [19]. Models like BioLORD-2023 enhance semantic representations of biomedical concepts, improving the accuracy of diagnoses and treatment recommendations [18]. In personalized medicine, LLMs identify genetic markers and predict treatment outcomes, enabling customized therapies based on genetic profiles [31].

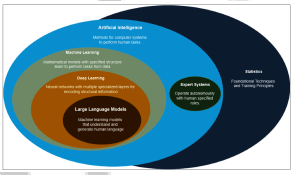
LLMs also enhance patient engagement by generating personalized health information and recommendations, crucial for managing chronic conditions [19]. They address challenges related to data interoperability, allowing seamless access to diverse health data sources for comprehensive patient profiling and informed treatment strategies [20]. Despite challenges like data privacy and model interpretability, ongoing advancements in LLM technology present significant opportunities for enhancing medical care quality and personalization [16].



(a) Large language model and downstream tasks in transcriptomics[27]



(b) Training and Output of a Machine Learning Model for Gene Expression Analysis[31]



(c) AI and its components[24]

Figure 4: Examples of Applications in Clinical and Personalized Medicine

As illustrated in Figure 4, integrating LLMs into bioinformatics transforms clinical and personalized medicine. The figures demonstrate LLM applications in transcriptomics, enhancing predictions of complex biological processes like RNA structures and protein expression. These models leverage extensive genomic data for intricate analyses, offering new insights into gene expression and regulatory mechanisms. The training and output of machine learning models for gene expression analysis highlight AI's potential in processing diverse biological inputs to produce clinically relevant outputs. The Venn diagram of AI components emphasizes the hierarchical relationship between AI, machine learning, and their applications, showcasing how these technologies converge to advance personalized medicine. Collectively, these examples underscore the pivotal role of LLMs and AI in revolutionizing bioinformatics, particularly in tailoring medical interventions to individual profiles [27, 31, 24].

4 Challenges and Limitations

The implementation of Large Language Models (LLMs) in bioinformatics faces several challenges that affect their efficacy and applicability. A significant issue is data quality and availability, which directly impacts LLM performance in healthcare and scientific research [32, 35, 26, 4, 12]. The accuracy of these models is inherently tied to the quality of training data, highlighting the necessity for comprehensive labeled datasets and an understanding of their implications for bioinformatics research.

4.1 Data Quality and Availability

The success of LLMs in bioinformatics is heavily reliant on the availability of high-quality data, essential for training models to produce accurate insights. A major hurdle is the need for extensive labeled datasets, which are crucial for fine-tuning LLMs for specific tasks within bioinformatics [8]. Creating and curating these datasets requires substantial time and resources.

Current benchmarks often lack systematic evaluation and tend to be confined to specific tasks, failing to fully exploit the potential of LLMs in bioinformatics [21]. This underscores the need for datasets

that capture the complexity and variability of biological data. Moreover, the computational challenges of managing large sequence datasets, such as those posed by the NP-hard RLCS problem, necessitate innovative data processing and model training solutions [15].

While frameworks like TourSynbio-Search improve access to bioinformatics resources, they may struggle with specialized queries requiring deep domain knowledge [11]. Complex queries can be further complicated by ambiguous user instructions, impeding LLM utilization in bioinformatics [22]. Addressing these challenges is crucial for enhancing LLM applications in bioinformatics, allowing researchers to tackle complex biological problems across various domains [27, 4, 21, 26].

4.2 Computational Resources and Efficiency

The deployment of LLMs in bioinformatics heavily depends on the computational resources required for effective training and implementation. Challenges such as the computational complexity of similarity calculations, exemplified by the SpanSeq method, highlight the difficulties of deploying LLMs in this field [9]. These challenges are compounded by the high computational demands of training LLMs on large datasets, necessitating high-performance computing resources like the Summit supercomputer and TPU Pods [14].

High computational costs limit the scalability and efficiency of LLMs in bioinformatics, affecting their feasibility in resource-constrained environments and hindering widespread adoption [8]. The reliance on advanced computational infrastructure raises concerns about the accessibility and sustainability of LLM-based solutions, especially for smaller research institutions with limited funding.

Addressing these computational challenges is essential for optimizing LLM efficiency in bioinformatics, requiring the development of more efficient algorithms and distributed computing frameworks to alleviate computational demands. Such improvements enhance speed and scalability, facilitating broader applications of these advanced AI systems across various fields [4, 12, 32, 26].

4.3 Model Interpretability and Transparency

Interpretability and transparency of LLMs are significant challenges in bioinformatics, where understanding model outputs is crucial for effective application. The black-box nature of LLMs complicates elucidating the reasoning behind predictions or classifications, which is essential for validating their use in sensitive domains like healthcare and genomics [17].

The computational expense associated with many LLMs limits their accessibility in bioinformatics, often confining their use to well-funded institutions with advanced infrastructure, raising concerns about the democratization of AI technologies [17].

To address these challenges, there is a pressing need for developing interpretable and transparent LLMs that provide clear insights into their decision-making processes. Models must not only be accurate but also capable of offering explanations accessible to both domain experts and non-experts. Enhancing interpretability is crucial for fostering trust and promoting widespread adoption of LLMs in bioinformatics, facilitating integration into biological research and healthcare applications. Establishing transparent evaluation frameworks will be essential for validating model capabilities, addressing limitations, and advancing healthcare outcomes through reliable AI applications [20, 21, 26, 27, 23].

4.4 Bias and Ethical Considerations

The application of LLMs in bioinformatics raises ethical concerns and potential biases that require careful examination. A critical challenge is the inherent biases in training data, which can significantly influence LLM outputs and affect their reliability and fairness [4]. These biases may stem from datasets used in training, potentially limiting result generalizability and raising ethical concerns in diverse biological contexts [14].

Ethical considerations also encompass aligning LLMs with human values to prevent generating harmful or misleading content [2]. The risk of biased or inaccurate outputs poses significant threats, especially in high-stakes applications like healthcare and genetic modeling, where errors can have severe consequences. This underscores the importance of robust assessment frameworks to evaluate the reliability and ethical implications of LLMs in bioinformatics [35].

Existing studies often inadequately address biases and the potential misuse of LLM-generated content in academic and professional contexts [32]. The expressiveness of ontologies and the validation of inferred knowledge within knowledge graphs are additional ethical considerations that must be addressed to ensure the integrity of bioinformatics research [33].

Moreover, reliance on the underlying quality of LLMs and domain-specific models (DSMs) can introduce variability based on specific models, complicating the ethical landscape of LLM applications [36]. Human oversight is essential to mitigate misinformation risks and ethical violations, ensuring responsible and effective use of LLMs in bioinformatics research [35].

4.5 Specialized Domain Knowledge and Adaptability

Deploying LLMs in bioinformatics requires a nuanced understanding of specialized domain knowledge to effectively address unique challenges posed by biological data. LLMs must adeptly incorporate domain-specific information to enhance predictive capabilities and interpretative accuracy in bioinformatics applications [16]. This is particularly critical given the complexity and variability inherent in biological systems, necessitating models that adapt to genomic, proteomic, and transcriptomic intricacies.

LLMs' adaptability to specialized knowledge is crucial for effective application in tasks such as protein structure prediction, where understanding evolutionary and functional protein contexts is essential [14]. They must also integrate diverse data types and sources, including genomic sequences, protein interactions, and clinical records, to provide comprehensive insights into biological phenomena [27].

The development of domain-specific LLMs, tailored for biomedical and clinical applications, underscores the importance of customizing models to meet the specific needs of various bioinformatics tasks [18]. These specialized models are designed to process and analyze complex biological data, enhancing the precision and relevance of their outputs in bioinformatics research.

Adapting LLMs to specialized domain knowledge involves addressing existing model limitations, such as their tendency to produce generic outputs for complex biological questions [16]. Developing frameworks that facilitate the integration of domain-specific knowledge into LLMs is essential for improving their adaptability and effectiveness in bioinformatics [17].

5 Future Directions and Opportunities

5.1 Advancements in LLM Architectures and Training Techniques

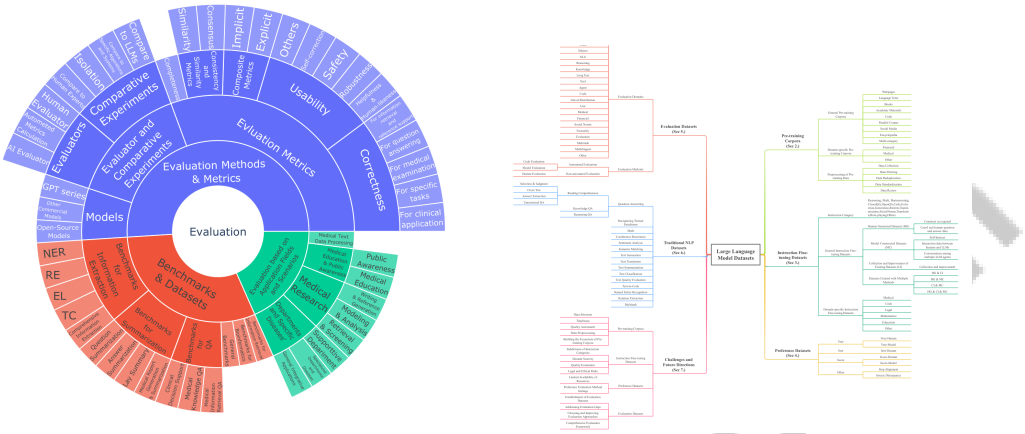
Advancements in Large Language Model (LLM) architectures and training techniques are pivotal for enhancing their application in bioinformatics. Future research should prioritize efficient training methodologies to overcome challenges related to model scalability and computational efficiency, which are essential for deploying LLMs effectively in bioinformatics [8]. This includes exploring hybrid models that merge the strengths of State Space Models (SSMs) and Transformers, potentially improving capabilities in in-context learning and multimodal understanding [7].

Investigating emergent abilities in LLMs and refining alignment techniques are critical for enhancing their utility in bioinformatics [2]. By focusing on these areas, researchers can boost LLM adaptability and performance, enabling better handling of the complex datasets typical of bioinformatics research.

Integrating LLMs with Domain-Specific Models (DSMs) offers further enhancement opportunities. This could involve incorporating additional modalities or refining calibration processes to improve the precision and relevance of LLM outputs in bioinformatics applications [36]. Additionally, employing multi-modal pre-training approaches may address current limitations of Transformer applications in bioinformatics, leading to more efficient models [17].

The ethical implications of deploying LLMs also warrant attention. Developing ethical AI frameworks, enhancing model transparency, and minimizing biases are essential for the responsible application of LLMs in bioinformatics [4]. This includes improving user training on responsible use and exploring LLM integration with other AI tools to enhance capabilities and ensure alignment with human values [32].

Specific model enhancements, such as those proposed for the PEvoLM model, could involve utilizing larger capacity LSTMs or extending training duration to improve performance in protein sequence analysis [13]. These advancements collectively aim to refine LLM architectures and training techniques, ultimately benefiting bioinformatics applications by enhancing efficiency, accuracy, and ethical alignment [1].



(a) AI Evaluation Methods and Metrics[26]

(b) Mind map categorizing topics related to natural language processing (NLP), including evaluation datasets and pre-training corpora[3]

Figure 5: Examples of Advancements in LLM Architectures and Training Techniques

As illustrated in Figure 5, significant advancements in LLM architectures and training techniques emerge as pivotal areas of focus. The first image encapsulates the multifaceted nature of AI evaluation methods and metrics, highlighting the importance of comprehensive evaluation frameworks encompassing models, benchmarks, datasets, and metrics. It emphasizes categories such as Named Entity Recognition (NER), Relation Extraction (RE), and Entity Linking (EL), essential for assessing AI model performance. The second image presents a mind map systematically organizing various topics related to NLP, categorizing essential components such as evaluation datasets, pre-training corpora, and instruction fine-tuning datasets. Together, these visual representations provide a cohesive overview of the current landscape and potential future advancements in LLM architectures and training techniques.

5.2 Integration with Other Modalities and AI Tools

Integrating Large Language Models (LLMs) with other AI tools and modalities presents a transformative opportunity to advance bioinformatics research. This synergy enhances LLM analytical capabilities by enabling the processing and interpretation of multimodal data, thereby providing a comprehensive understanding of complex biological systems [11]. By combining LLMs with modalities such as computer vision and network analysis, researchers can develop robust models that leverage diverse data types, including genomic sequences, protein structures, and clinical records, generating insights unattainable through single-modality approaches alone [18].

The integration of LLMs with domain-specific models (DSMs) and knowledge graphs enhances their ability to represent and reason about complex biological relationships. This approach facilitates the development of precise and contextually relevant models that can tackle specific bioinformatics challenges, such as predicting protein-protein interactions and identifying disease biomarkers [33]. Moreover, multimodal learning frameworks enable LLMs to incorporate visual and textual data, improving performance in tasks like protein structure prediction and functional annotation [17].

The potential for integrating LLMs with other AI tools extends to developing hybrid models that combine different architectures, such as Transformers and State Space Models (SSMs). These hybrid models can enhance the scalability and efficiency of LLMs, making them more suitable for processing large and complex datasets typical of bioinformatics research [7]. By leveraging the complementary

strengths of various AI tools, researchers can create powerful and versatile models that address unique bioinformatics challenges [36].

Furthermore, integrating LLMs with other AI modalities can facilitate the development of ethical and transparent models. By incorporating explainability techniques and ethical considerations into multimodal AI systems, researchers can ensure responsible and effective LLM use in bioinformatics research [4]. This approach enhances the trustworthiness of LLM outputs and aligns their applications with human values and ethical standards [32].

5.3 Dataset Expansion and Robust Evaluation

Benchmark	Size	Domain	Task Format	Metric
BioBench[21]	20,000	Bioinformatics	Named Entity Recognition	F1-score, Accuracy
LLM-lncRNA[30]	104,000	Transcriptional Regulation	Sequence Classification	MCC, F1 Score
SpanSeq[9]	19,171	Bioinformatics	Sequence Classification	MCC, Gorodkin
Gene-Benchmark[29]	312	Bioinformatics	Multi-label	AUC-ROC, F1
BioBench[37]	184	Bioinformatics	Programming Exercises	Success Rate, Attempts to Solve

Table 1: This table presents a comprehensive overview of various benchmarks used in bioinformatics for evaluating Large Language Models (LLMs). It details the size, domain, task format, and evaluation metrics of each benchmark, providing insights into their application and relevance in the field.

The advancement of Large Language Models (LLMs) in bioinformatics relies on expanding datasets and developing robust evaluation methodologies. Expanding datasets to encompass a broader range of biological data is crucial for improving the generalizability and applicability of LLMs across diverse biological contexts. Future research should focus on incorporating more diverse protein sequences into datasets, enhancing prediction accuracy and facilitating exploration of additional protein-related tasks [6]. Additionally, incorporating recent biomedical literature into model training can significantly enhance the knowledge base, further improving LLM capabilities [18].

Robust evaluation frameworks are vital for assessing LLM performance and reliability in bioinformatics applications. Table 1 provides a detailed overview of the benchmarks employed in bioinformatics to assess the performance and reliability of Large Language Models (LLMs), highlighting their size, domain, task format, and evaluation metrics. Developing new benchmarks tailored to bioinformatics tasks will enable more accurate assessments of LLM capabilities, highlighting areas for improvement and guiding future research directions [21]. These benchmarks should consider the unique characteristics of biological data and the specific challenges associated with analyzing complex omics datasets. Moreover, exploring alternative training strategies and refining self-reflection mechanisms can enhance model capabilities and adaptability to various bioinformatics tasks.

Optimizing methodologies such as the SpanSeq method exemplifies the need for further research into dataset expansion and robust evaluation, as these developments could significantly enhance applications to other biological data types and machine learning tasks. Enhancing ontology expressiveness and developing neuro-symbolic AI systems can improve knowledge acquisition and advance LLM integration in bioinformatics [33].

Improving model efficiency, addressing data imbalance, and exploring novel applications of foundation models in personalized medicine and drug discovery are also critical areas for future research [25]. By focusing on these aspects, researchers can ensure that LLMs effectively address the multifaceted challenges of biological research, ultimately contributing to advancements in life sciences and improved healthcare outcomes.

6 Conclusion

The exploration of Large Language Models (LLMs) in bioinformatics underscores their transformative impact, particularly in improving the precision and efficiency of biological data analysis. These models hold considerable promise in enhancing our comprehension of complex biological mechanisms, with notable applications in sequence analysis, protein structure prediction, and genomics. The effective application of LLMs in this domain relies on their adaptation to specific bioinformatics challenges, necessitating innovations in model design and integration with specialized datasets.

Advancements in LLM development, such as the incorporation of State Space Models (SSMs) for managing long sequences, demonstrate significant efficiency improvements. Nonetheless, to fully harness these advancements, further research is required to explore the potential of SSMs and other cutting-edge architectures across diverse bioinformatics applications. The establishment of comprehensive benchmarks and datasets is crucial, as they provide a solid foundation for training and evaluating LLMs, thus facilitating significant progress in biological research.

Despite these promising developments, several challenges remain, including the potential for bias and misinformation propagation. These issues necessitate careful consideration and the implementation of strategies to mitigate risks. The integration of LLMs into medical and bioinformatics practices must be accompanied by continuous refinement and ethical oversight to ensure their responsible and effective use.

References

[1] Saurabh Pahune and Manoj Chandrasekharan. Several categories of large language models (llms): A short survey. *arXiv preprint arXiv:2307.10188*, 2023.

[2] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[3] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*, 2024.

[4] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3, 2023.

[5] Marcin P. Joachimiak, Mark A. Miller, J. Harry Caufield, Ryan Ly, Nomi L. Harris, Andrew Tritt, Christopher J. Mungall, and Kristofer E. Bouchard. The artificial intelligence ontology: Llm-assisted construction of ai concept hierarchies, 2024.

[6] Bohdan B. Khomtchouk, Edmund Weitz, and Claes Wahlestedt. The machine that builds itself: How the strengths of lisp family languages facilitate building complex and flexible bioinformatic models, 2016.

[7] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges, 2024.

[8] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

[9] Alfred Ferrer Florensa, Jose Juan Almagro Armenteros, Henrik Nielsen, Frank Møller Aarestrup, and Philip Thomas Lanken Conradsen Clausen. Spanseq: Similarity-based sequence data splitting method for improved development and assessment of deep learning projects, 2024.

[10] Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. A fine-tuning dataset and benchmark for large language models for protein understanding, 2024.

[11] Yungeng Liu, Zan Chen, Yu Guang Wang, and Yiqing Shen. Toursynbio-search: A large language model driven agent framework for unified search method for protein engineering, 2024.

[12] Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, Xingtai Lv, Youbang Sun, Yang Li, Dong Li, Fuchu He, and Bowen Zhou. Automating exploratory proteomics research via language models, 2024.

[13] Issar Arab. Pevolm: Protein sequence evolutionary information language model, 2023.

[14] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021.

[15] Marko Djukanović, Jaume Reixach, Ana Nikolikj, Tome Eftimov, Aleksandar Kartelj, and Christian Blum. A learning search algorithm for the restricted longest common subsequence problem, 2024.

[16] Jinge Wang, Zien Cheng, Qiuming Yao, Li Liu, Dong Xu, and Gangqing Hu. Bioinformatics and biomedical informatics with chatgpt: Year one review, 2024.

[17] Nimisha Ghosh, Daniele Santoni, Indrajit Saha, and Giovanni Felici. A review on the applications of transformer-based language models for nucleotide sequence analysis, 2024.

[18] François Remy, Kris Demuynck, and Thomas Demeester. Biolord-2023: Semantic textual representations fusing llm and clinical knowledge graph insights, 2023.

[19] Maojun Sun. Llamacare: A large medical language model for enhancing healthcare knowledge sharing, 2024.

[20] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models, 2024.

[21] Hengchuang Yin, Zhonghui Gu, Fanhao Wang, Yiparemu Abuduhaibaier, Yanqiao Zhu, Xinming Tu, Xian-Sheng Hua, Xiao Luo, and Yizhou Sun. An evaluation of large language models in bioinformatics research, 2024.

[22] Joshua Pickard, Ram Prakash, Marc Andrew Choi, Natalie Oliven, Cooper Stansbury, Jillian Cwycyshyn, Alex Gorodetsky, Alvaro Velasquez, and Indika Rajapakse. Language model powered digital biology with brad, 2024.

[23] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

[24] Samuel A. Donkor, Matthew E. Walsh, and Alexander J. Titus. Computing in the life sciences: From early algorithms to modern ai, 2024.

[25] Qing Li, Zhihang Hu, Yixuan Wang, Lei Li, Yimin Fan, Irwin King, Le Song, and Yu Li. Progress and opportunities of foundation models in bioinformatics, 2024.

[26] Yining Huang, Keke Tang, Meilian Chen, and Boyuan Wang. A comprehensive survey on evaluating large language model applications in the medical industry, 2024.

[27] Jiajia Liu, Mengyuan Yang, Yankai Yu, Haixia Xu, Kang Li, and Xiaobo Zhou. Large language models in bioinformatics: applications and perspectives, 2024.

[28] Musa Nuri Ihtiyar and Arzucan Ozgur. Generative language models on nucleotide sequences of human genes, 2023.

[29] Yoav Kan-Tor, Michael Morris Danziger, Eden Zohar, Matan Ninio, and Yishai Shimoni. Does your model understand genes? a benchmark of gene properties for biological and text models, 2024.

[30] Wei Wang, Zhichao Hou, Xiaorui Liu, and Xinxia Peng. Exploring the potentials and challenges of using large language models for the analysis of transcriptional regulation of long non-coding rnas, 2024.

[31] Yanjun Lyu, Zihao Wu, Lu Zhang, Jing Zhang, Yiwei Li, Wei Ruan, Zhengliang Liu, Xiaowei Yu, Chao Cao, Tong Chen, Minheng Chen, Yan Zhuang, Xiang Li, Rongjie Liu, Chao Huang, Wentao Li, Tianming Liu, and Dajiang Zhu. Gp-gpt: Large language model for gene-phenotype mapping, 2024.

[32] Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil Van Der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023.

[33] Md. Rezaul Karim, Lina Molinas Comet, Md Shajalal, Oya Deniz Beyan, Dietrich Rebholz-Schuhmann, and Stefan Decker. From large language models to knowledge graphs for biomarker discovery in cancer, 2023.

[34] Chun-Ka Wong, Ali Choo, Eugene C. C. Cheng, Wing-Chun San, Kelvin Chak-Kong Cheng, Yee-Man Lau, Minqing Lin, Fei Li, Wei-Hao Liang, Song-Yan Liao, Kwong-Man Ng, Ivan Fan-Ngai Hung, Hung-Fat Tse, and Jason Wing-Hon Wong. Lomics: Generation of pathways and gene sets using large language models for transcriptomic analysis, 2024.

[35] Michael O'Neill and Mark Connor. Amplifying limitations, harms and risks of large language models, 2023.

[36] Tianyu Zhang, Yuxiang Ren, Chengbin Hou, Hairong Lv, and Xuegong Zhang. Molecular graph representation learning integrating large language models with domain-specific small models, 2024.

[37] Stephen R. Piccolo, Paul Denny, Andrew Luxton-Reilly, Samuel Payne, and Perry G. Ridge. Many bioinformatics programming tasks can be automated with chatgpt, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.