Received: 15 January 2025 | Accepted: 22 April 2025

DOI: 10.1002/qub2.70014

REVIEW ARTICLE

Quantitative Biology
Open Access

Large language models for bioinformatics

Wei Ruan[1] | Yanjun Lyu[2] | Jing Zhang[2] | Jiazhang Cai[3] | Peng Shu[1] | Yang Ge[4] | Yao Lu[4] | Shang Gao[5] | Yue Wang[1] | Peilong Wang[6] | Lin Zhao[1] | Tao Wang[3] | Yufang Liu[3] | Luyang Fang[3] | Ziyu Liu[3] | Zhengliang Liu[1] | Yiwei Li[1] | Zihao Wu[1] | Junhao Chen[1] | Hanqi Jiang[1] | Yi Pan[1] | Zhenyuan Yang[1] | Jingyuan Chen[6] | Shizhe Liang[7] | Wei Zhang[8] | Terry Ma[9] | Yuan Dou[10] | Jianli Zhang[10] | Xinyu Gong[10] | Qi Gan[10] | Yusong Zou[10] | Zebang Chen[10] | Yuanxin Qian[10] | Shuo Yu[10] | Jin Lu[1] | Kenan Song[10] | Xianqiao Wang[10] | Andrea Sikora[11] | Gang Li[12] | Xiang Li[13] | Quanzheng Li[13] | Yingfeng Wang[14] | Lu Zhang[15] | Yohannes Abate[16] | Lifang He[17] | Wenxuan Zhong[3] | Rongjie Liu[3] | Chao Huang[4] | Wei Liu[6] | Ye Shen[4] | Ping Ma[3] | Hongtu Zhu[5] | Yajun Yan[10] | Dajiang Zhu[2] | Tianming Liu[1]

[1]School of Computing, University of Georgia, Athens, Georgia, USA

[2]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas, USA

[3]Department of Statistics, University of Georgia, Athens, Georgia, USA

[4]Department of Epidemiology and Biostatistics, University of Georgia, Athens, Georgia, USA

[5]Department of Biostatistics, UNC Chapel Hill, Chapel Hill, North Carolina, USA

[6]Department of Radiation Oncology, Mayo Clinic, Phoenix, Arizona, USA

[7]Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, Georgia, USA

[8]School of Computer and Cyber Sciences, Augusta University, Augusta, Georgia, USA

[9]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[10]College of Engineering, University of Georgia, Athens, Georgia, USA

[11]Department of Biomedical Informatics, University of Colorado, Boulder, Colorado, USA

[12]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[13]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[14]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, Tennessee, USA

[15]Department of Computer Science, Indiana University Indianapolis, Indianapolis, Indiana, USA

[16]Department of Physics and Astronomy, University of Georgia, Athens, Georgia, USA

[17]Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania, USA

Correspondence
Tianming Liu and Dajiang Zhu.
Email: tliu@uga.edu and
dajiang.zhu@uta.edu

Abstract

With the rapid advancements in large language model technology and the emergence of bioinformatics-specific language models (BioLMs), there is a

Wei Ruan, Yanjun Lyu, and Jing Zhang are co-first authors.

growing need for a comprehensive analysis of the current landscape, computational characteristics, and diverse applications. This survey aims to address this need by providing a thorough review of BioLMs, focusing on their evolution, classification, and distinguishing features, alongside a detailed examination of training methodologies, datasets, and evaluation frameworks. We explore the wide-ranging applications of BioLMs in critical areas such as disease diagnosis, drug discovery, and vaccine development, highlighting their impact and transformative potential in bioinformatics. We identify key challenges and limitations inherent in BioLMs, including data privacy and security concerns, interpretability issues, biases in training data and model outputs, and domain adaptation complexities. Finally, we highlight emerging trends and future directions, offering valuable insights to guide researchers and clinicians toward advancing BioLMs for increasingly sophisticated biological and clinical applications.

KEYWORDS
bioinformatics-specific language models, biological systems, biomedical AI, large language models, life active factors

1 | INTRODUCTION

The rapid development of large language models (LLMs) such as BERT [1], GPT [2], and their specialized counterparts has revolutionized the field of natural language processing (NLP). Their ability to model context, interpret complex data patterns, and generate human-like responses has naturally extended their applicability to bioinformatics, where biological sequences often mirror the structure and complexity of human languages [3]. LLMs have been successfully applied across various bioinformatics domains, including genomics, proteomics, and drug discovery, offering insights that were previously unattainable through traditional computational methods [4].

Despite significant advancements, challenges remain in the systematic categorization and comprehensive evaluation of applications of these models on bioinformatics problems. Considering the variety of bioinformatics data and the complexity of life activities, navigating the field can often be challenging, as existing studies tend to focus on a limited scope of applications. This leaves gaps in understanding the broader utility of LLMs in various bioinformatics subfields [5].

This survey aims to address these challenges by providing a comprehensive overview of LLM applications in bioinformatics. By focusing on different levels of life activities, this article collected and exhibited related works from two major views: life science and biomedical applications.

We have collaborated with domain experts to compile a thorough analysis spanning key areas in these views, such as nucleoid analysis, protein structure and function prediction, genomics, drug discovery, and disease modeling, including applications in brain diseases and cancers, as well as vaccine development.

In addition, we propose the new term "Life Active Factors" (LAFs) to describe the molecular and cellular components that serve as candidates for life science research targets, which widely includes not only concrete entities (DNA, RNA, protein, genes, drugs) but also abstract components (bio-pathways, regulators, gene-networks, protein interactions) and biological measurements (phenotypes, disease biomarkers). LAFs is a comprehensive term that is capable of reconciling the conceptual divergence arising from research across various bioinformatics subfields, benefiting the understanding of multi-modality data for LAFs and their interplays in complex bio-systems. The introduction of LAFs aligns well with the spirit of foundational models and emphasizes the unification across sequence, structure, and function of the LAFs while respecting the interrelationships of each LAF as a node within the biological network.

By bridging existing knowledge gaps, this work seeks to equip bioinformaticians, biologists, clinicians, and computational researchers with an understanding of how LLMs can be effectively leveraged to tackle pressing problems in bioinformatics. Our survey not only highlights recent advances but also identifies open challenges and opportunities, laying the foundation for future interdisciplinary collaboration and innovation (Figure 1).

2 | BACKGROUND OF LANGUAGE MODELS AND FOUNDATION MODELS IN BIOINFORMATICS

Bioinformatics has become a fundamental and transformative field in life sciences, bridging computational techniques and biological research. It emphasizes the development and application of computational tools

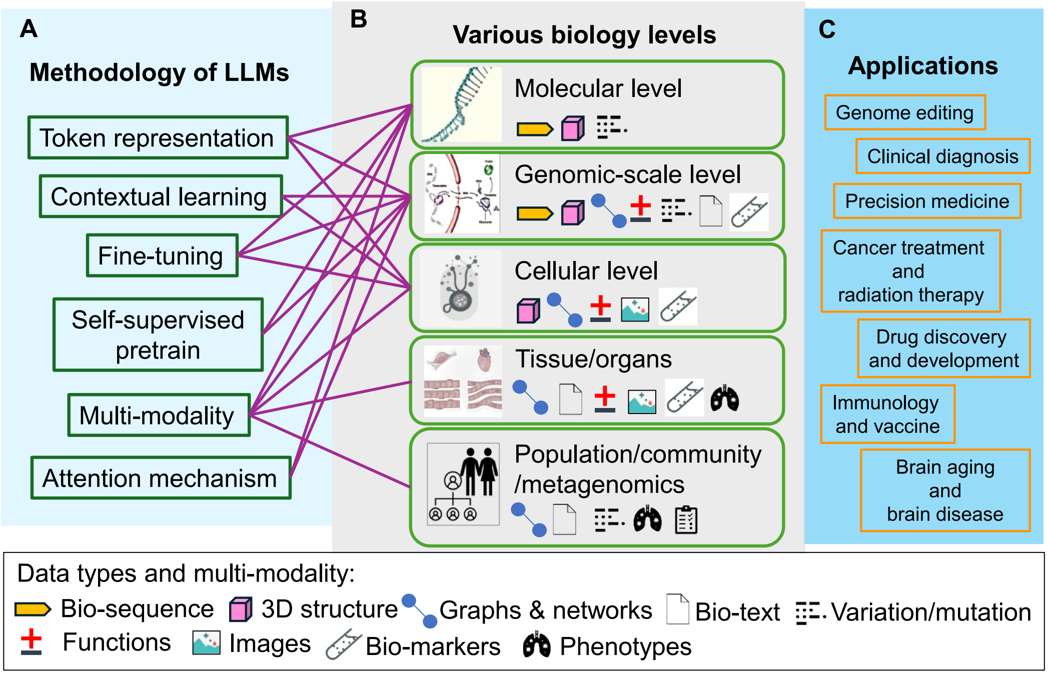

FIGURE 1 The applications of the methodology of large language models (LLMs) in bioinformatics tasks.

and methodologies to manage and interpret vast amounts of biomedical data, transforming them into actionable insights and driving advancements across diverse downstream applications. Modern computational tools, particularly those rooted in deep learning technology, have significantly accelerated the evolution of biological research.

The rapid advancements in LLMs technologies have inspired new approaches to bioinformatics computing. Considering the complexity of biological systems and highly structured nature of bioinformatics data, LLM-based computing methods have proven effective in addressing challenges across fields such as genomics, proteomics, and molecular biology. Inspired by LLM architectures like transformers, foundation models in bioinformatics excel at capturing complex patterns and relationships in biological data. They have evolved from single-modality tools to sophisticated multimodal systems, integrating diverse datasets such as genomic sequences and protein structures.

Central to their success is the availability of large-scale, high-quality training data and the adoption of self-supervised pre-training and fine-tuning techniques. These methods allow models to extract meaningful features from unlabeled data and adapt to specific bioinformatics tasks. Together with advances in architecture design, these innovations have broadened the capabilities and impact of foundation models, unlocking new insights into biological systems and accelerating progress in life sciences. The following sections discuss these advanced computing methods along with the intrinsic properties of biological systems and structured bioinformatics data.

2.1 | Foundations of language models and bioinformatics overview

2.1.1 | Basics of large language models and foundations models

Traditional language models are engineered to process and generate text in a human-like manner, leveraging the extensive datasets used during their training. These models excel at interpreting context, producing coherent and contextually appropriate responses, performing translations, summarizing text, and answering questions. LLMs are a type of foundation model trained on vast datasets to provide flexible and powerful capabilities [6–8] that address a broad spectrum of use cases and applications [8–78]. By efficiently handling diverse tasks, LLMs eliminate the need for building and training separate domain-specific models for each use case—a process that is often limited by cost and resource constraints. This unified approach not only fosters synergies across tasks but also frequently results in superior performance, making LLMs a more scalable and efficient solution. There are several key elements that make the language model successful in adaptation to bioinformatics tasks (Figure 1A).

Representation learning and tokenization

Tokenization in LLMs is influenced by the design of their tokenization algorithms, which primarily use subword-level vocabularies to represent text sequence data effectively. Popular tokenization algorithms, such as Byte-Pair Encoding (BPE) [79], WordPiece [80], and Unigram [81], are widely used. Although their vocabulary cannot perfectly capture every possible variation of input expressions, these tokenization methods effectively encode the features of words and their contextual relationships.

In the view of representation learning, the tokenization and token embedding algorithms of the language model generally succeeded in representing the hidden factors of variation behind the data. This representation is based on the unsupervised learning scheme of the language models. The sub-word context features learned in the encoder modules or embedding layers follow the probabilistic modeling and continuously update the representations on large corpus datasets [82].

Attention mechanism

LLMs widely use the transformer model [83, 84] as their foundational architecture. A core innovation of the transformer model is the multi-head self-attention mechanism, which establishes relationships among all relevant tokens, enabling more effective encoding of each word in the input sequence. The self-attention layer processes a sequence of tokens (analogous to words in a language) and learns context information across the entire sequence. The "multi-head" aspect refers to multiple attention heads operating simultaneously to capture diverse contextual features. Inside a single attention head, a token output embedding in a sequence is computed and fused with other tokens in the context with a proper causal mask. Such global level attention mechanic enables efficient information fusion along available context windows.

Self-supervised training methods

Language models are trained using self-supervised learning methods [85]. Unlike supervised learning, which typically requires human annotations, language models can leverage vast amounts of unannotated text data [86]. The objective of unsupervised learning is to analyze unlabeled data by identifying and capturing its meaningful properties. Neural networks can extend some of these approaches. For example, autoencoders compress data into a low-dimensional representation through a hidden layer known as the bottleneck layer and then reconstruct the original input data from this representation [87–90]. Language models leverage either the next word in a sentence as a natural label for the context or artificially mask a known word and predict it. This method, where unstructured data generates its own labels (e.g., predicting the next word or a masked word) and language models are trained to predict them, is known as self-supervised learning. Transformer-based models, with their parallel processing capabilities and ability to capture correlations across entire sequences, have achieved state-of-the-art (SOTA) performance [91, 92]. A more advanced training diagram is the text-to-text framework. This kind of training diagram unified multiple kinds of tasks, including translation, question answering (QA), classification, formulated and feeding to model as input and training it as a generative model to predict target text. This framework, which is named "T5" benefits using the same model, loss function, hyperparameters, etc. across a diverse set of tasks [93].

Pre-training methods

In many supervised learning problems, input data is represented by multiple features, comprising numerical or categorical information that can aid in making predictions. Scratch-trained models, which initialize and train all parameters from the ground up using task-specific datasets, typically require numerous iterations to converge fully on a single task. In general, transformer-based language models fall into two categories: scratch-trained models and pre-trained models. LLMs apply transformer-based pre-trained models that are trained from large amounts of unlabeled data and then fine-tuned for specific tasks. Pre-training learns general information from unlabeled data which can improve the convergence rate of the target tasks and often has better generalization than training parameters from scratch [94]. The use of context information in a large corpus to pre-train the whole model (or encoder modules) has achieved SOTA results in various downstream tasks.

2.1.2 | Bioinformatics applications and challenges

Using deep learning methods like language models to tackle bioinformatics problems is challenging. While deep learning models have shown superior accuracy in specific bioinformatics applications (e.g., genomics applications) compared to SOTA approaches and are adept at handling multimodal and highly heterogeneous data, significant challenges remain. Further work is required to integrate and analyze diverse datasets required for deep learning for genomic prediction and prognostic tasks. This is especially important for the development of explainable language models that can identify novel biomarkers and elucidate regulatory interactions across various biology levels: pathological conditions, including different tissues and disease states. These advancements require a deep understanding of complex bioinformatics data, the related tasks, and their mutual relationships [95]. In this review, we discuss such issues through two lenses: the various biology levels and the inherent regulations of life activities.

Various biology levels

Although no gold standard division was available, the levels of life-science factors in bioinformatics can be divided into five levels, from micro to macro. Here, we take the mammal model organisms as a template, the levels can be divided into: the molecular level, the genome-scale level, the cellular level, the tissue/organ system level, and the population/community/metagenomics level (Figure 1B). Bioinformatics often focuses on the first three levels (i.e., the molecular level, the genomic-scale level, and the cellular level). The molecular level analysis targets biologically active molecules, which include nucleic acids, amino acids, and other small bioactive molecules, and the relative experiments aimed at interpreting the life activities at this scale. The genomic-scale level models life activities from DNA, RNA, and proteins to metabolomics. The most famous regulation at the genomic scale level is The Central Dogma, which reveals the intrinsic relations of main life-activity factors on a sub-cellular scale. The whole sub-cellular system is modeled hierarchically, beginning with DNA, messenger RNA (mRNA), and proteins, extending to metabolomics, and ultimately inferring the phenotype [96]. At the cellular level, understanding cellular mechanisms is a fundamental challenge in biology and holds significant importance in biomedical fields, particularly concerning disease phenotypes and precision medicine. Using genes (the specific sequences of nucleotides within DNA that control downstream life activities) as a unit, the functions of genes and the gene products are essential research targets at this level. A comprehensive, structured, computation-accessible representation of gene function and variations is crucial for bioinformatics understanding of the cellular organism or virus. At the same time, the gene networks and mutual influences of gene products pose a challenge for such areas. Single cell sequencing technologies allow us to obtain gene expression data at the mRNA level, providing a foundation for analyzing entire cellular systems. This data is now extensively used to identify cell states during development, characterize specific tissues or organs, and evaluate patient-specific drug responses. In this review, the molecular components at the genomics level and cellular levels and their respective sets are collectively referred to as LAFs. It is important to note that the sequence representation format is the most commonly observed for each LAF. However, multi-modality data for LAF is also significant for representing the property of LAF, that is, the highly structured data format to record the function descriptions, abundancy, variations, and expressions [97, 98].

Data bias and risks

One key challenge in integrating LLMs into biomedical workflows is training data bias, where demographic imbalances or inadequate representation of minority populations can yield models that perpetuate or amplify existing health disparities. Additionally, clinical settings demand high reliability, given that any mistake or "hallucination" in a recommendation can have severe consequences for patient care. Ensuring privacy and confidentiality poses another difficulty: models must adhere to data protection standards such as Health Insurance Portability and Accountability Act (HIPAA) or general data protection regulation (GDPR) while still leveraging large-scale patient information [99, 100]. Furthermore, interpretability and transparency remain central concerns, as both healthcare professionals and regulatory bodies must be able to understand and scrutinize model-driven outcomes. In response, researchers and practitioners have explored several methods to address these issues. Bias mitigation techniques, including balanced sampling strategies and thorough bias audits, aim to correct or minimize skewed model outputs. Robust model validation, such as stress-testing on multiple, diverse datasets, helps reveal systematic errors and instill greater confidence in performance. To maintain patient privacy, teams often employ differential privacy, federated learning (FL), or secure multi-party computation, ensuring that sensitive data are either well-anonymized or kept on premises. Finally, explainable AI tools and rigorous human-in-the-loop approaches are increasingly adopted, providing healthcare professionals with interpretable reasoning traces while allowing them to override model suggestions when necessary. By combining these strategies, the biomedical community can work toward LLM solutions that are both technically sound and ethically responsible.

Inherent regulations of life activities

Since most LAFs at each biological level are represented in a sequence format, transformer-based pre-trained language models are particularly well-suited for analyzing these sequences. An emerging consensus suggests that these sequences embody an underlying language that can be deciphered using language models. However, to play the roles in life activities, an essential logic of a single LAF is "sequences—structures—functions." Take proteomics analysis as an example, protein sequences can be viewed as a concatenation of letters from the amino acids, analogously to human languages. The latest protein language models utilize these formatted letter representations of secondary structural elements, which combine to form domains responsible for specific functions. The protein language models also direct inference of full atomic-level protein structure from primary sequence and produce functional proteins that evolution would require hundreds of millions of years to uncover [101–103].

In life activities, there are important regulation relationships among the LAFs across different levels as well as intra-level relationships. Considering the genomics level, genes control hereditary traits primarily by regulating the production of RNA and protein products. According to the central dogma of molecular biology, genes within DNA are transcribed into mRNA, which is then translated into gene products, such as proteins. For any given gene product, whether RNA or protein, its origin can be traced back to the gene that directed its synthesis. This traceability highlights that fully understanding a gene's functionality requires considering not only the gene itself but also the roles and functions of all its associated products. Genes regulate each other and create feedback loops to form cyclic chains of dependencies in gene regulatory networks, graph neural network-styled operations are suitable to model the "steady state" of genes. It is the same for proteins in protein-protein interactions (PPI). In the layer of pathways, it is a hypergraph where each hyperedge is a pathway including multiple proteins.

Within the cellular level, pathways integrate individual genes or protein products to perform certain cell functions under mutual intra-level regulations. Proteins interact with one another in various ways, such as inhibiting, activating, or combining with others, thereby influencing expression levels or protein abundances within cells. These interactions are collectively referred to as PPI. Some databases systematically organize results by annotating functionalities using Gene Ontology, utilizing the unique gene identifiers assigned to each gene within the genomic system [104, 105].

Scalability and performance limitations

Despite their potential, current LLM methods still face notable hurdles in handling extensive and heterogeneous bioinformatics datasets. Training or fine-tuning these models typically demands substantial computational resources and specialized hardware, which can become cost-prohibitive for large-scale omics or clinical data. Memory constraints also limit the maximum input length and model size, leading to trade-offs between complexity and efficiency. In addition, performance often deteriorates when LLMs are directly applied to highly domain-specific data without sufficient fine-tuning or domain adaptation. As a result, researchers are exploring more efficient architectures, model compression techniques, and scalable distributed training methods to ensure LLMs remain feasible and robust for the growing demands of biomedical data analysis [106, 107].

Model robustness

It is also one of the key challenges in applying LLMs to bioinformatics and requires careful attention. To enhance reliability and mitigate risks such as overfitting, misuse, and erroneous predictions in critical medical scenarios, several strategies can be implemented. Regularization techniques, such as dropout and weight decay, help improve generalization, while adversarial training enhances resilience against input variations. Uncertainty quantification methods, including Monte Carlo dropout and confidence calibration, can provide more reliable predictions. Additionally, fine-tuning on high-quality biomedical datasets and incorporating continual learning frameworks improve domain adaptability. Model interpretability techniques, such as attention visualization and SHapley Additive exPlanations, further enhance transparency and trust. Addressing biases in training data through fairness-aware training and human-in-the-loop validation ensures ethical and responsible deployment. Strengthening model robustness through these approaches is essential for the safe and effective application of LLMs in bioinformatics and healthcare [108, 109].

Continual learning

With the continuous generation of new biomedical data each year, many models are still trained on outdated datasets, making them less effective in capturing recent advancements. To integrate new data without retraining from scratch, continual learning has become a crucial technique for enhancing the application of LLMs in bioinformatics. Continual learning enables LLMs to dynamically incorporate newly published biomedical literature, updated genomic databases, and evolving clinical guidelines while retaining previously acquired knowledge. Techniques such as replay-based methods, regularization strategies, and dynamic architecture adjustments help mitigate catastrophic forgetting, ensuring long-term learning stability [110, 111]. In bioinformatics, continual learning not only improves the predictive capabilities of LLMs but also enhances their adaptability in biomolecular sequence analysis, personalized medicine, and drug discovery. By adopting continual learning frameworks, LLMs can remain up-to-date, accurate, and efficient for real-world biomedical research and healthcare applications [112, 113].

Privacy protection

In bioinformatics, protecting sensitive medical and genomic data while leveraging the power of LLMs presents a significant challenge, especially when different healthcare institutions possess proprietary datasets that cannot be publicly shared. FL offers a promising solution by enabling collaborative model training across multiple institutions without exposing raw data. In an FL framework, each institution trains a local instance of the model on its own data, and only the model updates (e.g., gradients or parameters) are shared with a central aggregator, ensuring data privacy. This decentralized approach not only enhances security and compliance with regulations such as HIPAA and GDPR but also allows LLMs to benefit from diverse, real-world biomedical data without compromising confidentiality. Recent advancements in privacy-preserving techniques, such as differential privacy and secure multi-party computation, further strengthen the security of FL in biomedical applications. By adopting FL, LLMs can be trained on distributed, sensitive datasets from different institutions, improving their generalizability and robustness while maintaining strict data privacy standards [114–116].

Efficient LLMs

As LLMs become increasingly complex, their high computational costs pose challenges for bioinformatics applications [117]. Model distillation helps address this by transferring knowledge from a large model (teacher) to a smaller one (student), reducing size while maintaining performance. This approach has been applied in tasks such as protein structure prediction and biomedical text mining [118]. Pruning techniques, which remove redundant parameters, further enhance efficiency by reducing memory and computational demands [119]. Recent advances in sparse LLMs and quantization make it possible to deploy these models in resource limited environments, enabling faster and more efficient bioinformatics analyses.

2.2 | Training methods and models

2.2.1 | Pre-training methods

Pre-training is a critical phase in the development of LLMs, where a model learns foundational linguistic representations by training on extensive and diverse datasets. This process typically employs self-supervised learning techniques such as masked language modeling (e.g., BERT [1]) or causal language modeling (e.g., GPT [2]), enabling the model to predict masked tokens or the next word in a sequence. Unlike traditional deep neural networks (DNNs) [120], which are often pre-trained on domain-specific datasets such as ImageNet [121], pre-training of LLMs is conducted on significantly larger datasets comprising diverse domains, including books, encyclopedias, and web content. Moreover, pre-training LLMs involves models with billions or even trillions of parameters, making it computationally and resource-intensive compared to conventional DNNs.

The primary advantage of pre-training lies in the model's ability to generalize across diverse language tasks, often achieving zero-shot [122] or few-shot [123] performance without additional task-specific training. This broad generalization enables LLMs to excel in tasks spanning natural language understanding, generation, and reasoning. However, the disadvantages of pre-training include high computational and energy costs, often requiring distributed systems with high-performance hardware. Additionally, pre-trained models can inherit biases and errors present in the training corpus [124], potentially leading to biased or undesirable outputs.

2.2.2 | Fine-tuning methods

Fine-tuning is the subsequent stage that builds upon the pre-trained model by adapting it to specific tasks or domains through additional supervised or semi-supervised training. This process utilizes smaller, targeted datasets and optimizes the model for a specific use case. Fine-tuning can be categorized into task-specific fine-tuning [125, 126], where models are specialized for particular tasks such as sentiment analysis or machine translation; domain-specific fine-tuning [11, 38], which refines the model for specialized fields such as medicine or law; and instruction fine-tuning [62, 127], where the model is trained to respond to natural language prompts in an aligned manner. Recent advancements in parameter-efficient fine-tuning methods [128], such as LoRA (low-rank adaptation) [129] and adapters [130], have further improved the efficiency of this process by updating only a subset of the model's parameters while maintaining the computational benefits of the pre-trained foundation.

Fine-tuning enhances the model's performance on specific tasks by leveraging domain- or task-specific data, achieving SOTA results in various applications. However, it introduces challenges such as the risk of overfitting to the fine-tuning dataset, potentially diminishing the model's generalization capabilities. Furthermore, fine-tuning requires high-quality labeled data to ensure reliability and accuracy in specialized applications.

2.2.3 | Reinforcement learning with human feedback methods

RALF [131] represents a crucial additional stage in the training pipeline of LLMs, designed to align model outputs with human preferences and expectations. While pre-training and fine-tuning equip the model with general linguistic understanding and task-specific expertise, RALF optimizes the model's behavior to produce responses that are more aligned with human values, instructions, or conversational styles, which is particularly critical for applications such as conversational agents, where user interaction quality is paramount.

RALF involves three primary components: a reward model trained on human-labeled preferences, a reinforcement learning (RL) algorithm to optimize the model's behavior based on the reward model, and iterative human feedback to refine the reward system. The reward model is typically developed by collecting a dataset of model outputs ranked by human evaluators. This ranking serves as the ground truth to train the reward model, which predicts the desirability of a given output. Subsequently, RL algorithms, such as proximal policy optimization [132], adjust the model parameters to maximize the reward score predicted by the reward model. Besides, direct preference optimization [133] algorithms operate on a dataset of ranked preferences, directly optimizing the model to prefer the more highly ranked output in each pair.

The primary advantage of RALF is its ability to align the outputs of a pre-trained and fine-tuned model with human expectations, improving qualities such as coherence, relevance, and ethical compliance. This approach is particularly effective in mitigating undesirable behaviors, such as generating toxic, biased, or irrelevant content. Furthermore, RALF enables the incorporation of domain-specific human expertise, allowing models to better serve niche applications. However, RALF introduces several challenges. First, the quality of human feedback is critical, poorly designed feedback mechanisms or misaligned human preferences can lead to suboptimal or even harmful model behavior. Second, RALF requires significant resources for human annotation and computationally expensive RL training. Furthermore, over-optimization for the reward model can lead to undesirable artifacts, such as the model exploiting weaknesses in the reward system rather than genuinely improving its outputs—a phenomenon known as "reward hacking" [134, 135].

2.2.4 | Knowledge distillation

Knowledge distillation (KD) has emerged as a key approach for efficient training and deploying LLMs by transferring the knowledge embedded in high-capacity teacher models to smaller, more efficient student models [136]. In essence, the student model learns to mimic both the predictive outcomes and the internal representation patterns of the teacher, thereby significantly reducing computational costs and memory demands during the pre-training phase [137–139]. This methodology promotes the development of leaner LLMs without sacrificing their ability to perform complex language tasks.

Recent advancements in KD have extended beyond final output matching. Modern methods utilize established LLMs to generate not only predictions but also detailed reasoning steps, which are often referred to as chain-of-thought sequences or intermediate logic traces [140, 141]. These rich annotations can then be incorporated into the fine-tuning process, enabling the target LLM to acquire deeper problem-solving skills and enhance interpretability without extensive manual labeling. By integrating these reasoning pathways, KD no longer serves solely as a compression mechanism but also imparts advanced critical thinking and inference capabilities to newly trained models. Moreover, recent work explores expanding KD to support specialized or domain-specific tasks where the established teacher models can guide the target LLM toward focusing on task-relevant knowledge, filtering out less pertinent information [142, 143]. This approach helps produce models that are better aligned with their intended applications. Additionally, a Bayesian perspective on KD has been introduced, offering a transparent interpretation of its statistical foundations and equipping the target model with robust uncertainty quantification capabilities [144, 145].

The integration of pre-training, fine-tuning, KD, and RALF represents a comprehensive training paradigm for LLMs. Pre-training serves as the foundation, equipping the model with general knowledge and linguistic capabilities through large-scale unsupervised learning. Fine-tuning adapts the model to specific tasks or domains, enhancing its performance in targeted applications. KD supports efficiency by enabling the transfer of knowledge from established teacher models to target models, while RALF refines the model's behavior to align with human preferences, ensuring outputs are both functionally accurate and socially acceptable. These stages are complementary and iterative. Insights gained during RALF can inform improvements in fine-tuning datasets or methodologies, while advancements in fine-tuning and KD can enhance the quality of RALF outcomes. Together, this pipeline not only ensures that LLMs are powerful and versatile but also makes them more usable and aligned with human-centered goals. This multi-stage training paradigm has been instrumental in the development of SOTA models like OpenAI's ChatGPT and Anthropic's Claude, setting a benchmark for future advancements in the field. These advancements include the release of both full-scale and lightweight versions, with KD often playing a role in optimizing the latter [146].

2.3 | Bioinformatics-specific datasets

The rapid advancements in LLMs have significantly propelled the development of bioinformatics by enabling more efficient data interpretation and knowledge extraction. LLMs excel in understanding, processing, and generating complex textual and numerical data, making them powerful tools for tasks such as sequence analysis, annotation, and predictive modeling [147, 148]. Leveraging bioinformatics-specific datasets, LLMs can further refine their understanding to address domain-specific challenges, transforming raw data into tangible, interpretable forms that accelerate research and innovation.

Currently, there are several publicly available datasets and benchmarks designed to train and evaluate LLMs in the bioinformatics domain. For instance, repositories such as PDBbind and BindingDB facilitate testing of protein-ligand prediction tasks, while UniProt and Pfam provide comprehensive protein sequence and functional information [149, 150]. Genomic databases such as dbSNP enable evaluation of variant interpretation pipelines, and specialized corpora—including BioASQ or CORD-19—support assessments in literature mining and biomedical QA [151]. Leveraging these domain-specific datasets helps ensure that LLMs not only perform well on general tasks, but also meet the specialized demands of large-scale bioinformatics applications.

2.3.1 | Question answering dataset

QA systems play a vital role in biomedicine, assisting with clinical decision support and powering medical chatbots. The development of robust QA systems relies heavily on diverse and well-curated datasets. Over the past decade, several biomedical QA datasets have been introduced, each targeting specific challenges and domains. For instance, MedMCQA [152] and MedQA [153] focus on general medical knowledge, providing open-domain questions and multiple-choice answers derived from medical licensing and entrance exams. GeneTuring targets genomics-specific tasks, such as gene name conversion and nucleotide sequence alignment. Meanwhile, BioASQ [154, 155] and PubMedQA [156] incorporate supporting materials, such as PubMed articles, to answer domain-specific questions with formats ranging from yes/no to multi-class classifications. These datasets are crucial for benchmarking QA systems, as they provide domain-specific contexts and evaluation metrics that drive the development of more accurate and reliable models tailored to biomedical needs.

2.3.2 | Text summarization dataset

Text summarization (TS) in biomedical and healthcare is a critical application of NLP, enabling the condensation of complex medical texts into concise, informative summaries without compromising essential details. This task is particularly valuable in areas such as the summarization of literature, the summarization of radiology reports, and the summarization of clinical notes. Among these, the summarization of radiology reports plays an essential role in transforming detailed imaging reports—including X-rays, CT scans, MRI scans, and ultrasounds—into easily understandable summaries. Datasets such as MIMIC-CXR [157] are instrumental in advancing this field, providing a large-scale resource with 473,057 chest X-ray images and 206,563 corresponding reports. Such data sets are essential for training and evaluating summarization models, offering domain-specific content and structured formats that drive improvements in both accuracy and reliability, ultimately enhancing clinical workflows and decision making.

2.3.3 | Information extraction dataset

Information extraction (IE) in biomedicine involves organizing unstructured text into structured formats through tasks such as named entity recognition (NER) and relation extraction (RE). Robust IE systems rely on high-quality datasets for training and evaluation. For instance: datasets such as BC5CDR [158], NCBI-disease [159], ChemProt [160–162], DDI [163], GAD [164], BC2GM [165], and JNLPBA [166] have become benchmarks for NER and RE tasks, addressing challenges involving diseases, chemicals, genes, and other biomedical entities. These datasets are essential benchmarks for tackling real-world biomedical challenges, enabling the development of more accurate and generalizable models.

LLMs have also shown potential in various biomedical tasks such as coreference resolution and text classification. The effectiveness of these applications often depends on the availability of high-quality datasets. For coreference resolution, datasets such as MEDSTRACT [167], FlySlip [168], GENIA-MedCo [169], DrugNerAR [170], BioNLP-ST'11 COREF [171], HANAPIN [172] and CRAFT-CR [173] provides essential benchmarks for identifying links between mentions of the same entity in biomedical texts. Pre-trained models such as BioBERT [174] and Span-BERT [175] have achieved notable success in this domain. In text classification, datasets such as HoC (comprising 1580 manually annotated PubMed abstracts for multi-label classification of cancer hallmarks) [176] have been pivotal.

In summary, the rapid progress in LLMs have transformed biomedical applications by improving data interpretation, knowledge extraction, and task automation. From QA and TS to IE, LLMs have demonstrated their potential across a wide range of bioinformatics-specific tasks. Central to their success is the availability of high-quality, domain-specific datasets, which are indispensable for training, benchmarking, and refining these models to address real-world challenges. These datasets not only enhance the effectiveness of LLMs but also act as a driving force in advancing the field of bioinformatics and biomedicine. As the availability of diverse and richly annotated datasets continues to expand, they will fuel the integration of LLMs into increasingly complex and specialized applications. Looking to the future, combining bioinformatics-specific datasets with cutting-edge techniques promises to unlock groundbreaking solutions, enabling more precise, efficient, and scalable innovations that will shape the next generation of biomedical research and healthcare.

2.4 | Model evolution and key milestones

The evolution of LLMs in bioinformatics has marked a transformative journey. Initially developed for NLP tasks, these models, such as BERT [1] and GPT [177], have demonstrated remarkable potential in addressing challenges specific to the bioinformatics domain. Leveraging their ability to process and generate sequences, LLMs have been adapted for various biological data types, including DNA, RNA, proteins, and drug molecules [3].

In genomics, models such as DNABERT [178] and GROVER [179] are trained on DNA sequences to predict functional regions, such as promoters and enhancers, and analyze mutations. Similarly, transcriptomics benefits from models such as SpliceBERT [180] and RNA-FM [181], which assist in understanding RNA splicing and secondary structure prediction. For proteomics, PPLMs such as ProtTrans [182] and ProtGPT2 [183] enhance predictions related to protein structure, function, and interactions. These advances are made possible by the foundational transformer architecture, which excels at processing sequential data. Fine-tuning these pre-trained models for domain-specific tasks extends their utility to applications in drug discovery, where simplified molecular input line entry system (SMILES) representations of molecules and protein sequences are integrated to predict interactions and properties.

A notable breakthrough in bioinformatics has been the AlphaFold series, which has applied cutting-edge machine learning to solve protein structure prediction challenges. AlphaFold2 (AF2) revolutionized structural biology with its unprecedented accuracy in predicting protein structures based solely on amino acid sequences. Its attention-based deep learning architecture captured intricate protein folding patterns, surpassing traditional physics-based and homology-modeling methods. By leveraging evolutionary information through multiple sequence alignments (MSAs), AF2 provided reliable predictions even in the absence of experimental data, significantly reducing the time and costs associated with obtaining protein structural information, accelerating advancements in drug discovery and functional genomics [184].

Building on AF2's success, AlphaFold3 (AF3) introduced groundbreaking capabilities, particularly in modeling protein complexes, including protein-peptide interactions. Transitioning from individual protein structure predictions to multi-component biological assemblies, AF3 addressed challenges protein-protein docking and protein-peptide interaction modeling. Through its template-based and template-free approaches, further extended the versatility and impact of the AlphaFold series [185].

Key features of AlphaFold3 enhanced accuracy in complex structures: AF3 excels in predicting protein-peptide complex structures, achieving a high percentage of accurate models in challenging scenarios; innovative template-free modeling: while maintaining strengths in TB predictions, AF3 introduces powerful template-free algorithms that allow for diverse model generation with reliable accuracy, even in the absence of homologous structural data; and sophisticated scoring and ranking: AF3 integrates advanced scoring metrics such as DockQ and MolProbity, ensuring accurate evaluation of predicted structures. Its models show fewer issues such as twisted peptides or cis non-proline residues, reflecting improved protein-like properties and geometric quality.

The progression from AF2 to AF3 reflects the iterative refinement of computational methods to address increasingly complex biological problems. While AF2 focused on individual protein structures, AF3 emphasizes dynamic interactions within biological systems, signaling a shift toward a more holistic understanding of molecular biology. These innovations underscore how machine learning continues to redefine bioinformatics, enabling accurate and efficient modeling of protein structures and interactions. The AlphaFold series exemplifies the potential for transformative breakthroughs in biology and medicine, paving the way for future applications in understanding complex biological systems.

3 | APPLICATIONS FOR BIOINFORMATICS PROBLEMS

At the heart of LLMs lies the transformer architecture, which leverages an attention mechanism to manage word importance in context without the traditional constraints of recurrent (RNN) or convolutional (CNN) neural networks. The self-attention mechanism of transformers not only allows for robust parallelization and scalability but also excels at capturing long-range dependencies in text. In bioinformatics, the growing availability of extensive datasets across diverse tissues, species, and modalities presents both an opportunity and a challenge. Bioinformatics analysis typically seeks to uncover hidden relationships within vast amounts of data, which can be broadly categorized into two formats: molecular and cellular. Molecular data often consist of sequences—strings of four bases for DNA and RNA, and strings of 20 different amino acids for proteins. Cellular data, such as that from single-cell RNA-seq, single-cell ATAC-seq, or single-cell CITE-seq, typically takes the form of a count matrix with cells as rows and modalities as columns. While there are parallels between these data types and the structured data used in NLP, significant differences pose unique challenges for applying LLMs directly.

A comprehensive LLM framework for bioinformatics involves three critical stages: data tokenization, model pre-training, and subsequent analyses. Due to the inherent differences between bioinformatics and conventional NLP data, researchers have been pioneering adaptations to the LLM architecture to better suit bioinformatics applications. The following section will provide a detailed overview of notable contributions in this evolving field.

3.1 | Genome level

Genome data primarily provide molecular-level insights, focusing on the sequences of DNA and RNA. This format bears a strong resemblance to natural language, as it is structured as ordered sequences of strings. In this analogy, each nucleotide in a sequence read is akin to a character, each read is akin to a sentence, and the entire genome is comparable to the full article. To bridge the genome sequence and natural language, multiple studies try several ways to tokenize the genome sequence to make it similar to the concept of "word" in the natural language. To gain deeper insight into the functionalities of various genome segments, most studies apply the BERT (bidirectional encoder representations from transformers) as the core model, which excels in understanding the functions of a genome segment in relation to its surrounding genome region and is easily extended to different specific tasks by fine-tuning the model with specific dataset.

3.1.1 | LLM for DNA analysis

In DNA analyses, biological sequences are encoded into structured tokens to facilitate effective model processing. A commonly adopted method involves tokenizing sequences into *k*-mers, typically ranging from 3 to 6 bases in length. This approach creates a vocabulary of *k*-mer permutations analogous to words in natural language, allowing the pre-trained model to decipher patterns within these *k*-mers. The choice of *k* directly affects the complexity and size of the resulting library, presenting a trade-off between modeling efficiency and accuracy.

One of the pioneering methods, DNABERT [178], tokenizes DNA sequence data using overlapping fixed-length *k*-mers, as well as the recently developed Nucleotide Transformer [186]. To enhance model efficiency, subsequent versions such as DNABERT-2 [187] and GROVER [179] have employed BPE [79], a statistical compression technique that iteratively merges the most frequently co-occurring genome segments. This method extends beyond fixed *k*-mer lengths, significantly improving the efficiency and generalizability of the models. HyenaDNA [188] uses one-mer to tokenize the DNA sequence since it uses Hyena [189] as the core model, which allows much longer input than BERT. Additionally, some models integrate supplementary data into their tokenization process; for instance, DNAGPT [190] incorporates species information, and MuLan-Methyl [191] combines sequence and taxonomy data into a natural language-like sentence to fully leverage existing LLM capabilities.

In terms of pre-training approaches, many models utilize the BERT architecture with a masked learning method for self-supervised training. To boost training efficiency, DNABERT incorporates the AdamW optimizer with fixed weight decay and applies dropout to the output layer. DNABERT-2 introduces enhancements such as attention with linear biases (ALiBi) [192] and flash attention [193]. In contrast, the MuLan-Methyl framework integrates five fine-tuned language models (BERT and four variants) for the joint identification of DNA methylation sites, maintaining consistency with their original pre-training setups. DNABERT-S [194] develops a contrastive learning-based method to help effectively cluster and separate different species. Some methods adopt other LLM models. For example, DNAGPT uses a GPT-based model and the next-token prediction for its pre-training, enabling it to forecast subsequent tokens based on previous ones. HyenaDNA uses Hyena, a new LLM model that allows a longer context input, to study long-range genomic sequence properties.

When applying these models to specific bioinformatics tasks, most integrate additional task-relevant data for fine-tuning. For instance, DNABERT and its derivatives utilize the Eukaryotic Promoter Database (EPDnew) [195] to predict gene promoters, the ENCODE database [196] for transcription factor binding site identification, and dbSNP for functional variant detection. MuLan-Methyl uses data from three main types of DNA methylation across multiple genomes for accurate predictions. Nucleotide Transformer includes multiple downstream tasks by fine-tuning the model with different datasets, such as using histone ChIP-seq data [196] for epigenetic marks prediction, using human enhancer elements data [197] for enhancer sequence prediction, and using human annotated splice sites data [198] for splice site prediction. DNAGPT leverages data on polyadenylation signals and translation initiation sites for genomic signal and region recognition. Moreover, due to the generative nature of GPT, DNAGPT can also generate artificial human genomes without additional fine-tuning data. Without further fine-tuning, some methods use the embedding from the model directly. DNABERT-S can be used for species clustering and classification.

3.1.2 | LLM for RNA analysis

Unlike DNA, RNA analysis encompasses more complex and varied tasks, requiring tailored pre-processing strategies. RNABERT [199], mirroring the structure of DNABERT, employs the *k*-means method for tokenizing RNA sequences. Given the typically shorter sequences of RNA compared to DNA, other models such as SpliceBERT [200], RNA-MSM [201], and RNA-FM [181] utilize single nucleotides for tokenization. In addition to sequence tokenization, these models often incorporate metadata during preprocessing. For instance, RNA-RBP [202] labels each sequence as positive or negative based on the presence of an RNA-binding protein (RBP) region, while SpliceBERT similarly labels sequences for RNA-splicing sites. RNA-MSM enhances its input by including MSAs [203] to preserve the evolutionary history of sequences.

The pre-training approach for RNA largely follows that of DNA, utilizing BERT's architecture and masked language modeling for training. Specifically, RNA-MSM adopts a structure akin to AlphaFold2 [204], leveraging an MSA-transformer architecture. Depending on the target application, models are pre-trained with different datasets: RNABERT and RNA-MSM use sequences from the Rfam database, RNA-FM utilizes non-coding RNA sequences from RNAcentral [205], and SpliceBERT is pre-trained with RNA sequences from 72 vertebrates available on the UCSC Genome Browser [206]. BERT-RBP is trained using the eCLIP-seq dataset, which includes RBP information [207].

Once trained, the BERT-based models process tokenized sequences to produce embeddings for each token. These embeddings are directly utilized in several applications; RNABERT employs them to classify RNAs from different families, while BERT-RBP uses them to predict RBP-binding sites. Furthermore, the attention maps generated as part of the model output play a critical role: SpliceBERT uses these maps to assess the impact of genetic variants on RNA splicing, BERT-RBP to analyze transcript region types and predict secondary structures, and RNA-MSM for secondary structure and solvent accessibility predictions.

For task-specific enhancements, some models undergo fine-tuning with additional datasets. SpliceBERT, for example, is fine-tuned using a human Branchpoints dataset [208] to predict BP sites and the Spliceator dataset [209] to assess splice sites across species. RNA-FM is fine-tuned with the PDB dataset [210] to facilitate RNA 3D structure reconstruction.

3.2 | Gene products level

With advances in single-cell technologies, researchers have gained enhanced insights into the functional roles and regulatory mechanisms of gene products within individual cells [211]. Single-cell RNA sequencing (scRNA-seq) data, which records the expression levels of various genes across individual cells, is particularly instrumental. Typically presented in a count matrix format, scRNA-seq data contrasts with sequence data; it lacks a natural order and contains numerical values rather than sequences of strings. Researchers have explored various methods to adapt this data for compatibility with LLMs, adjusting the representation of scRNA-seq data to harness the power of LLM methodologies.

To adapt scRNA-seq data for LLM compatibility, researchers have devised various strategies. Models such as Cell2Sentence [212], tGPT [213], and Geneformer [214] employ a ranked sequence of gene symbols by expression level as inputs. ScGPT [215] and scBERT [216] discretize gene expressions and treat them as tokens. Additionally, scGPT incorporates metadata for position embedding, while scBERT leverages gene2vec [217] to capture semantic similarities based on general co-expression.

Some methods utilize transformer-based architecture, which accommodates non-discrete inputs more flexibly. CIForm [218] segments the gene expression vector of each cell into equal-length sub-vectors or patches. TOCICA [219] groups gene expression into patches representing specific pathways, and ScTransSort [220] employs CNNs to generate gene-embedding patches, transforming the expression matrix into multiple 2D square patches. TransCluster [221] uses linear discriminant analysis to convert gene expression counts into embedding vectors.

Unlike genome analyses, single-cell analyses adopt diverse model architectures for pre-training. For instance, Cell2Sentence, tGPT, and scGPT utilize GPT, whereas scBERT and Geneformer are based on BERT architecture. Transformer-based methods often integrate a linear classifier post-transformer and train a supervised model using cell types, as seen in CIForm, TOCICA, scTransSort, and TransCluster.

The primary aim of scRNA LLM methodologies is to achieve accurate and generalized cell type annotations across various tissues and species. Supervised transformer-based methods use the pre-trained model directly for cell-type annotation. For instance, tGPT supports developmental lineage inference, and TOCICA enables interpretable dynamic trajectory analysis. LLM-based methods, post-pre-training, can be fine-tuned for specialized tasks or data-scarce scenarios. ScGPT is adaptable for tasks such as cell annotation, perturbation response prediction, batch effect correction, and gene regulatory network inference. Similarly, Geneformer can be fine-tuned to predict gene dosage sensitivity, chromatin dynamics, and gene network dynamics.

3.3 | Epigenomics

Decoding the information residing in the non-coding portion of the genome is one of the fundamental challenges in genomics [222]. While substantial progress has been made in understanding the coding regions of the genome, non-coding regions remain poorly understood, particularly their roles in disrupting the regulatory syntax of DNA and their contributions to gene regulation. Existing LLMs, for example, Enformer [223], which take DNA sequences as input and perform downstream tasks, face two critical limitations: they cannot predict the functions of sequences in different cellular contexts, and they fail to incorporate 3D chromatin interaction data.

EpiGePT [224] is a new LLM designed to overcome these challenges. It enables researchers to predict functionality in diverse cellular contexts and integrate 3D chromatin interaction data into genomic modeling. EpiGePT's architecture consists of four key components: a sequence module that analyzes DNA sequences, a transcription factor module that encodes cellular contexts, a transformer module that examines long-range interactions between DNA regions, and a prediction module that outputs context-specific gene regulation insights. To predict function in novel cellular contexts, EpiGePT employs its TF module, which represents the expression and binding activities of hundreds of transcription factors as a context-specific vector. This vector is then combined with DNA sequence features, which are tokenized into genomic bins—each representing a segment of the DNA sequence. These tokens, enriched with both sequence and context-specific TF features, form the input to the model, ensuring it captures both the local sequence information and the cellular context. This approach allows the model to treat each genomic bin as a token with embedded positional and biological context, leveraging the self-attention mechanism in the transformer module to learn long-range interactions and context-specific functionality. EpiGePT also addresses the challenge of incorporating 3D chromatin interaction data, which is critical for understanding long-range gene regulation. It guides the self-attention mechanism of its transformer module using ground truth 3D interaction data, such as HiChIP [225] or Hi-C [226] loops. This alignment is achieved through a cosine similarity loss that adjusts the attention weights to reflect known 3D genomic interactions. By doing so, EpiGePT can model regulatory mechanisms, such as enhancer-promoter interactions, with higher fidelity than existing models.

3.4 | Protein level

Mass spectrometry (MS)-based proteomics focuses on characterizing proteins within complex biological samples [227, 228]. Recent advancements in MS technology have enabled researchers to generate vast amounts of proteomics data [229]. However, the rapid growth in data volume presents significant analytical challenges. To tackle these issues, Ding et al. introduced PROTEUS, an LLM-based tool designed for automating proteomics data analysis and hypothesis generation [230]. PROTEUS leverages a foundational LLM to integrate and coordinate existing bioinformatics tools, facilitating scientific discovery from raw proteomics data. Protein sequences share many similarities with natural language, and since breakthroughs have been achieved in applying NLP methods to protein sequence research, a variety of protein language models have emerged, differing in architecture, training strategies, and application scope [4, 231, 232]. Here, we outline the main types of protein language models and downstream tasks, each tailored to address distinct bioinformatics challenges in protein modeling, structure prediction, and functional annotation.

3.4.1 | Models for protein LLM

Encoder-only models

Encoder-only models, such as BERT-based models primarily designed for understanding protein sequences. These models excel in tasks that involve recognizing patterns within the sequences, making them suitable for protein classification, mutation effect prediction, and secondary structure analysis. Examples include ESM 1b [233], ESM-1v [234], ProteinBert [235], ProtTrans [236], which leverage the bidirectional attention mechanisms of BERT to capture contextual relationships within amino acid sequences.

Decoder-only models

Decoder-only models, such as the GPT family in NLP, focus on generating new sequences based on learned distributions. In protein research, these models can be applied to generate synthetic protein sequences with desired properties or to design novel proteins. Models such as ProGen [237], ProtGPT2 [238], ZymCTRL [239], RITA [240], IgLM [241], ProGen2 [242], and PoET [243] are notable for their ability to produce diverse protein sequences that exhibit specific biochemical functions. This category is instrumental in protein engineering and synthetic biology, where the generation of novel, functional proteins is crucial [231].

Encoder-decoder models

Encoder-decoder models combine the strengths of both encoder-only and decoder-only architectures, making them highly adaptable to a range of protein-related tasks. They are particularly effective for sequence-to-sequence tasks, such as protein sequence alignment, where aligning amino acid sequences accurately is essential for understanding evolutionary relationships. These models can be fine-tuned for protein structure prediction or protein-protein interaction mapping, contributing to advancements in fields such as drug discovery and disease diagnosis. The models include Fold2Seq [244], MSA2Prot [245], Sgarbossaetal [246], Leeetal [247], LM-Design [248], MSA-Augmenter [249], ProstT5 [250], xTrimoPGLM [251], SS-pLM [252], pAbT5 [253], ESM-GearNet-INR-MC [254].

Multi-modal protein models

Multi-modal protein models integrate traditional protein language models with additional data types, such as structural and interaction information, to create powerful frameworks capable of analyzing both sequence and structural features simultaneously. By integrating textual protein sequences with structural annotations, these models enhance predictive capabilities for tasks such as 3D protein structure prediction, binding interaction analysis, and functional site identification. Frameworks such as multimodal protein representation learning (MPRL) [255] exemplify this approach by combining sequence information, 3D structural data, and functional annotations to capture the complex characteristics of proteins. For example, MPRL employs evolutionary scale modeling [256] for sequence analysis, variational graph autoencoders for residue-level graphs, and PointNet autoencoders for 3D point cloud representations. This comprehensive data integration preserves both spatial and evolutionary aspects of proteins, allowing the model to generalize effectively across tasks such as protein–ligand binding affinity prediction and protein fold classification. Similarly, Models such as ProtTrans [236] and ESM [256] treat protein sequences as textual data, to learn rich embeddings that, when combined with 3D structural data, improve predictions of structure-function relationships. This multimodal synergy is essential for advancing protein engineering and drug discovery, mapping complex biological functions onto computational representations of proteins.

3.4.2 | Downstream tasks for protein LLM

Protein modeling, especially through deep learning approaches, addresses a variety of critical tasks in biological research and medicine. For instance, deep learning methods are extensively applied in PPIs, which are fundamental for cellular functions [257]. This prediction aids in understanding disease mechanisms, drug-target interactions, and the structural features of proteins that contribute to complex molecular pathways. The prediction of PPIs also enables the identification of novel therapeutic targets, providing significant insights for drug discovery and design. The typical models include AlphaFold [258], AlphaFold 2 [259], AlphaFold 3 [185], Graph-BERT [260], MARPPI [261].

Large-scale models also excel in predicting protein post-translational modifications (PTMs), which play essential roles in regulating protein function, stability, and cellular signaling [262]. Various machine learning models, including those based on transformers and neural networks, have been adapted to predict PTM sites with improved accuracy. For instance, the PTMGPT2 model [263], developed by fine-tuning a GPT-2 architecture, leverages prompt-based approaches to identify subtle sequence motifs that correspond to PTM sites across diverse types [264]. By using custom tokens in its prompt, PTMGPT2 effectively captures sequence context and improves prediction accuracy, making it useful for identifying disease-associated mutations and potential drug targets.

Additionally, protein structure prediction remains a pivotal task in computational biology. It involves understanding how proteins fold and how their structures determine functions. Advanced models, such as those using transformer architectures, facilitate the accurate prediction of protein structures, providing crucial information for synthetic biology, enzyme design, and therapeutic protein engineering [265]. These methods enable scientists to predict protein folding patterns and design novel proteins with specific functions, potentially revolutionizing fields such as drug discovery and synthetic biology. The typical models include AlphaFold [258], AlphaFold 2 [259], AlphaFold 3 [185], ColabFold [266], Eigenfold [267].

The development of protein large language models (Prot-LLMs) relies on diverse datasets that capture the complexity of protein sequences and functions. These datasets typically include unlabeled data for unsupervised pre-training, such as protein sequences from repositories such as UniProt [268], AlphaFoldDB [269] which houses millions of protein sequences across species. For fine-tuning and evaluation, labeled datasets focus on specific protein characteristics, such as structure, function, and interactions. Examples include datasets for secondary structure prediction, protein-protein interaction networks, and specific PTM sites [270]. These labeled datasets enable Prot-LLMs to perform tasks such as function annotation, PTM prediction, and protein structure modeling.

3.5 | Metabolomics

Metabolomics represents the comprehensive analysis of the complete set of small-molecule metabolites within a biological system, providing a snapshot of the cellular biochemical status at a given time. This omics discipline is pivotal in elucidating the dynamic interactions between genotype and phenotype, as metabolites are the end-products of cellular processes and are directly involved in the regulation of biological functions. Metabolomics has emerged as a powerful tool in various areas of biological and medical research, including the identification of biomarkers for disease diagnosis, prognosis, and therapeutic monitoring [3], as well as the elucidation of molecular mechanisms underlying disease pathogenesis. The integration of LLMs into metabolomics offers transformative potential for analyzing and interpreting metabolomic data. With their capacity to process vast amounts of textual and numerical information, LLMs, particularly transformer-based models adapted for biological data, have shown promise in metabolite identification and pathway analysis.

3.5.1 | Data integration and interpretation

One of the most significant challenges in metabolomics is the integration and interpretation of large, complex datasets. LLMs can facilitate the integration of metabolomic data with other omics data (e.g., genomics, transcriptomics, proteomics) and clinical data, a challenge increasingly addressed by dynamic modeling approaches to enhance our understanding of metabolic phenotypes [271]. By processing and analyzing these multi-omics datasets, LLMs can identify patterns and correlations that may not be apparent through traditional statistical methods. For instance, LLMs can be trained to predict the biological pathways and processes associated with specific metabolite profiles, thereby providing insights into the molecular mechanisms of disease.

Recent advances in multi-modal LLM architecture have addressed key challenges in data integration. The development of cross-attention mechanisms specifically designed for metabolomic data has improved the ability to handle heterogeneous data types. These mechanisms allow for simultaneous processing of spectral data, chemical structures, and biological annotations. However, significant challenges remain in handling the high dimensionality and sparsity of metabolomic data. Novel approaches incorporating dimensionality reduction techniques and attention-based feature selection have shown promise in managing these challenges while maintaining biological relevance.

3.5.2 | Biomarker discovery and validation

The identification of robust biomarkers is a critical aspect of metabolomics, with applications in disease diagnosis, prognosis, and therapeutic monitoring. LLMs can be employed to analyze large datasets from clinical trials and cohort studies to identify potential biomarkers associated with specific disease states. Integrated deep learning frameworks have addressed challenges such as matching uncertainty and metabolite identification, enabling more reliable biomarker discovery and validation through the integration of diverse data sources [272]. This can lead to the development of more accurate and reliable biomarker panels for clinical use.

The validation of metabolomic biomarkers presents unique challenges that LLMs are increasingly equipped to address. Recent developments in uncertainty quantification for LLMs have improved the reliability of biomarker predictions. Statistical frameworks incorporating false discovery rate control and multiple hypothesis testing have been integrated into LLM-based biomarker discovery pipelines. Furthermore, the development of interpretable deep learning architectures has enhanced our ability to understand the biological mechanisms underlying identified biomarkers, leading to more robust validation processes.

3.5.3 | Metabolic pathway analysis and drug discovery

Metabolomics data can provide valuable insights into the perturbations of metabolic pathways in disease states. LLMs exhibit remarkable capabilities in analyzing biological data, such as genomic sequences and protein structures, making them instrumental in identifying druggable targets and novel therapeutic compounds [5]. For example, LLMs can be trained to predict the effects of gene variants on enzyme activity and metabolic fluxes, thereby aiding in the identification of druggable targets. Additionally, LLMs can be used in the discovery of novel therapeutic compounds by predicting the binding affinity of small molecules to metabolic enzymes and pathways.

Advanced graph neural network architectures have emerged as powerful tools for metabolic pathway analysis when integrated with LLMs. These hybrid approaches can capture both the topological structure of metabolic networks and the chemical properties of individual metabolites. Recent developments in attention-based graph neural networks have improved our ability to predict metabolic flux distributions and identify regulatory bottlenecks. The integration of molecular docking simulations with LLM-based predictions has enhanced the accuracy of drug-target interaction (DTI) predictions in metabolic pathways.

3.5.4 | Personalized medicine

The application of metabolomics in personalized medicine is rapidly gaining momentum, with the potential to tailor treatments to individual patients based on their metabolic profiles. LLMs can play a crucial role in this context by analyzing patient-specific metabolomic data in conjunction with genomic, proteomic, and clinical data to develop personalized treatment plans. For instance, LLMs can be used to predict the response of individual patients to specific therapies based on their metabolic profiles, thereby enabling the selection of the most effective treatment options.

3.5.5 | Literature mining and knowledge discovery

The vast amount of published literature in the field of metabolomics presents both an opportunity and a challenge for researchers. LLMs can be employed to mine this literature for relevant information, such as the identification of novel metabolites, the characterization of metabolic pathways, and the discovery of new biomarkers, addressing the challenge of synthesizing metabolomics research [273]. By processing and analyzing textual data from scientific articles, LLMs can generate hypotheses and identify trends that may guide future research directions.

3.5.6 | Quality control and data standardization

The reproducibility and comparability of metabolomics data are critical for the advancement of the field. Tools such as the LargeMetabo package facilitate the reproducibility and standardization of large-scale metabolomics datasets, ensuring consistency across studies. LLMs can be used to standardize metabolomics data by identifying and correcting inconsistencies in data annotation, nomenclature, and reporting. Additionally, LLMs can assist in the development of quality control metrics and standards for metabolomics experiments, thereby improving the reliability and comparability of metabolomics data across different studies and platforms.

3.5.7 | Predictive modelling and simulation

LLMs can be integrated with machine learning models to develop predictive models of metabolic pathways and networks. Advanced multivariate models, including machine learning techniques, have shown efficacy in analyzing metabolomics data to uncover predictive patterns of metabolic pathways [274]. These models can be used to simulate the effects of genetic, environmental, and pharmacological perturbations on metabolic processes, thereby providing insights into the molecular mechanisms of disease and the potential outcomes of therapeutic interventions. Furthermore, LLMs can be used to predict the outcomes of metabolic engineering strategies in synthetic biology applications, such as the optimization of metabolic pathways for the production of biofuels, pharmaceuticals, and other valuable chemicals.

The integration of LLMs into metabolomics represents a significant advancement in the field, with the potential to enhance data analysis, interpretation, and knowledge discovery. By leveraging the power of LLMs, researchers can unlock the full potential of metabolomics data, leading to new insights into disease mechanisms, the development of novel therapeutic strategies, and the advancement of personalized medicine. As LLMs continue to evolve, their applications in metabolomics are expected to expand further accelerating the pace of discovery and innovation in this exciting field.

4 | DISEASE-SPECIFIC BIO-MEDICAL APPLICATIONS

The application of LLM technology to medical-related bioinformatics data offers significant potential to enhance various downstream biomedical tasks (Figure 1C).

4.1 | Brain ageing and brain disease

LLMs are transforming the study and management of brain diseases by enabling innovative approaches to diagnosis, treatment, and knowledge discovery. These models excel in processing diverse data types including clinical notes, imaging studies, biological sequences, and brain signals, unlocking new possibilities for identifying disease patterns, predicting progression, and personalizing care. This section highlights the diverse applications of LLMs in brain diseases, focusing on three critical areas: clinical diagnostic support, therapeutic assistance, and information driven decision-making. Through these contributions, LLMs address longstanding challenges in managing complex neurological conditions, offering scalable and non-invasive solutions that enhance both research and clinical practice.

4.1.1 | Clinical diagnosis support

Accurate and timely diagnosis is the foundation of effective medical care, particularly in complex and progressive conditions such as neurodegenerative diseases. The emergence of LLMs in healthcare offers transformative potential in clinical diagnostics by leveraging their advanced capabilities in processing diverse forms of unstructured data. From textual data to biological sequences and brain signals, LLMs excel at identifying patterns, extracting clinically relevant information, and supporting decision-making. Additionally, their ability to integrate multimodal data has shown promise in improving diagnostic accuracy. This section explores how LLMs are applied to various data types crossing different brain diseases, highlighting their unique advantages and current challenges in clinical diagnosis.

Textual data—Biomedical text

LLMs are increasingly applied to the analysis of biomedical textual data, including literature and electronic health records (EHRs). This form of biomedical textual data closely mirrors the fundamental structure of LLMs. LLMs can identify significant insights within medical reports, enhancing diagnostic accuracy. In brain disease research, LLMs have been leveraged to diagnose conditions such as seizures, Alzheimer's disease (AD), headaches, strokes, Parkinson's disease (PD), and other neurodegenerative disorders using textual data from clinical notes, MRI reports, and neuropathological records. For AD, LLMs provide a non-invasive, cost-effective, and scalable solution by analyzing unstructured data within EHRs. For example, Mao et al. demonstrated that the LLM can accurately predict mild cognitive impairment (MCI) to AD progression using clinical notes as the early detection [275]. Feng et al. utilized LLMs to embed textual data in alignment with imaging data, significantly enhancing AD diagnosis through a multimodal approach [276]. Beyond AD, LLMs have also shown promise in managing epilepsy, with studies successfully classifying seizure-free patients and extracting seizure frequency and other critical information from clinical notes [277]. Additionally, in a study analyzing neurodegenerative disorders at the Mayo Clinic, diagnostic accuracies of 76%, 84%, and 76% were achieved using ChatGPT-3.5, ChatGPT-4, and Google Bard, respectively, underscoring the potential of LLMs in generating differential diagnoses for complex neuropathological cases [278]. EHRs also include detailed MRI reports, which are critical in neurological diagnoses. Bastien Le Guellec et al. evaluated the performance of LLMs in extracting information from real-world emergency MRI reports, demonstrating high accuracy without requiring additional training [279]. Similarly, Kanazawa et al. showed that a fine-tuned LLM could classify MRI reports such as no brain tumor, post-treatment brain tumor, and pre-treatment brain tumor with accuracy comparable to human readers [280]. These results highlight the growing importance of LLMs in processing MRI reports, which are essential components of EHRs, further enhancing their utility in brain disease diagnosis and management.

Textual data—Transcription text

In addition to text-based data, transcriptions from speech data are increasingly valuable for diagnosing brain diseases that impair linguistic abilities. Patients with AD, for example, often exhibit distinct speech patterns when describing images, including word-finding difficulties, grammatical errors, repetitive language, and incoherent narratives. The ADReSS Challenge dataset inspired the research community to develop automated methods to analyze speech, acoustic, and linguistic patterns in individuals to detect cognitive changes, frequently used in such studies [281–284]. LLMs outperform traditional methods such as support vector machine, and random forest in this context. The existing work also shows that the combination of acoustic features with linguistic features for a multi-model can improve the performance. The maximum accuracy obtained by the acoustic feature is 64.5%, and the BERT model provides a classification accuracy of 79.1% over the test dataset, the fusion of the acoustic model with the BERT model shows an improvement of 6.1% classification accuracy over the BERT model [282]. Linguistic analysis is also pivotal in diagnosing aphasia, a disorder commonly caused by left-hemisphere strokes. Chong et al. evaluated the clinical efficacy of LLM surprisal in a study where post-stroke aphasia patients narrated the story of Cinderella after reviewing a wordless picture book. The approach revealed significant potential for quantifying deficits and improving aphasia discourse assessment [285].

Textual data—Text generation

In addition to biomedical text and speech data, recent advancements in text generation have further showcased the potential of LLMs in clinical applications. Studies indicate that LLM-generated summaries are often preferred over those produced by human experts across various domains, including radiology reports, patient inquiries, progress notes, and doctor-patient dialogues [286]. This demonstrates the capacity of LLMs to synthesize complex clinical information effectively. Techniques such as Chain-of-Thought (CoT) prompting and text classification have been introduced to improve the confidence and precision of LLM outputs. For example, when applied to neurologic cases, GPT-4 has shown promising results. By analyzing history and neurologic physical examination (H&P) data from acute stroke cases, GPT-4 accurately localized lesions to specific brain regions and identified their size and number. This was achieved through Zero-Shot Chain-of-Thought and text classification prompting, highlighting the model's potential for advanced neuroanatomical reasoning [287]. Similarly, in AD diagnostics, prompting LLMs with clinical Chain-of-Thought frameworks has enabled them to generate detailed diagnostic rationales, demonstrating their ability to support reasoning-aware diagnostic frameworks [288].

Biological sequences

The process of DNA transcription to RNA, followed by translation into proteins, is fundamental to life and is often referred to as The Central Dogma of molecular biology. Many brain diseases, including AD, PD, autism spectrum disorder, and frontotemporal dementia, are closely associated with abnormalities in DNA, RNA, or protein sequences. To investigate the genetic and molecular mechanisms underlying these diseases, approaches such as genome-wide association studies, transcriptome analysis, and proteomic profiling have been widely utilized. However, traditional methods often struggle to interpret the complex patterns present in these large-scale datasets. LLMs, with their advanced capabilities in processing sequential data, offer a transformative approach for analyzing biological sequences, enabling deeper insights into disease mechanisms and potential therapeutic targets. Several innovative LLMs have been developed for biological sequences. For DNA, models such as Enformer [223], Nucleotide Transformer [186], and DNABERT [178] have shown significant promise. For RNA, RNABERT [289], RNAFM [181], and RNA-MSM [181] focus on structural inference and functional predictions. For proteins, models such as ProteinBERT [235], ESM-1b [233], and ProtST [290] have demonstrated capabilities in understanding sequence-function relationships. Despite these advances, the application of LLMs to reveal relationships between abnormalities in biological sequences and specific brain diseases remains limited. Notable exceptions include epiBrainLLM, proposed by Liu et al., which extracts genomic features from personal DNA sequences using a retained LLM framework and combines these features to enhance diagnosis [291]. This approach provides valuable insights into the causal pathways linking genotypes to brain measures and AD-related phenotypes. Another study utilized LLMs to predict protein phase transitions such as amyloid aggregation, a key pathological feature of age-related diseases such as AD, demonstrating the potential of LLMs in advancing molecular-level understanding of neurodegenerative disorders [292].

Brain signal

Brain signal data, including sMRI, fMRI, and electroencephalogram (EEG), is critical for diagnosing and understanding various brain diseases. Abnormalities in these signals are key diagnostic indicators for conditions such as epilepsy, attention-deficit/hyperactivity disorder (ADHD), and mental health disorders. For epilepsy, EEG abnormalities such as seizures, spikes, and slowing patterns are widely used for diagnosis. A fine-tuned LLM, named EEG-GPT, was developed for classifying EEG signals as normal or abnormal, showing strong performance in identifying these patterns [293]. Similarly, Liu et al. leveraged LLMs to guide affinity learning for rs-fMRI, enabling comprehensive brain function representation and improved diagnostic accuracy for brain diseases [294]. All the LLM models above are based on the transformer architecture. Due to the long-range dependencies and temporal resolution in brain signals, Mamba-based LLM also shows its potential in this field. Behrouz and Hashemi proposed BrainMamba, an efficient encoder for modeling spatio-temporal dependencies in multivariate brain signals. It combines a time-series encoder for brain signals and a graph encoder for spatial relationships, making it versatile for neuroimaging data. With a selective state space model design, BrainMamba achieves linear time complexity, enabling training on large-scale datasets. Evaluations on seven real datasets across three modalities (fMRI, magnetoencephalography (MEG), EEG) and tasks such as seizure, ADHD, and mental state detection show that BrainMamba outperforms baselines with lower time and memory requirements [295].

4.1.2 | Therapeutic assistance

LLMs have demonstrated a strong capability to engage in conversations on daily life topics, personal matters, and specific concerns. When fine-tuned to provide empathetic and understanding responses, they hold significant potential as tools for companionship and emotional support. This capability is particularly valuable for individuals with dementia (PwD), who often experience social isolation. Research indicates that social isolation is strongly linked to an increased risk of developing dementia later in life [296]. Addressing social isolation plays a vital role in mitigating cognitive decline among the elderly. Recent studies have explored the potential of LLMs to alleviate social isolation and provide therapeutic support. For example, Qi demonstrated that ChatGPT effectively reduces feelings of loneliness among older adults with MCI by offering conversational engagement and cognitive stimulation [297]. Similarly, Raile highlighted the dual role of ChatGPT as a complement to psychotherapy and an accessible entry point for individuals with mental health concerns who have yet to seek professional help [298]. These findings suggest that LLMs can serve as valuable tools to support mental health and cognitive functioning in vulnerable populations.

In the context of neurodegenerative diseases, wearable devices integrated with AI technologies offer promising avenues for continuous monitoring and personalized care. Mohammed and Venkataraman introduced an AI-powered wearable device that leveraging LLMs to monitor the daily activities of patients with PD by analyzing multimodal data such as tremors, movements, and posture [299]. This approach enables real-time and personalized assessments of disease progression, potentially enhancing patient care and quality of life.

Language impairments, such as aphasia, present significant challenges in communication. Binta Manir et al. utilized BERT models to predict and complete sentences for individuals with aphasia, thereby improving the accuracy of speech prediction [300]. This approach benefits caregivers and speech therapists by facilitating more effective communication strategies and supporting rehabilitation efforts.

Brain-computer interfaces (BCIs) further exemplify the integration of advanced AI techniques into healthcare. Over recent decades, BCIs have provided novel solutions for various neurodegenerative disorders, including AD [301] and PD [302]. The incorporation of advanced AI algorithms, such as machine learning and deep learning, has significantly enhanced BCI performance, improving neuroergonomic systems, human-robot interactions, and robotic-assisted surgeries [303, 304]. Notably, integrating LLMs with BCIs introduces unique opportunities, such as reliably comprehending users' emotional states to create emotionally aware conversational agents [305] and decoding attempted speech from the brain activity of paralyzed patients [306]. These advancements highlight the transformative potential of LLMs in facilitating communication and enhancing the quality of life for individuals with severe disabilities.

Collectively, these studies underscore the versatile applications of LLMs as therapeutic assistance in brain diseases. By enhancing social interaction, providing cognitive support, enabling continuous monitoring, and assisting in communication, LLMs represent a promising avenue for improving patient outcomes and overall quality of life.

4.1.3 | Information driven decision-making

LLMs have proven to be valuable tools for information retrieval, serving as vast repositories of knowledge. Reza Saeidnia et al. reported that dementia caregivers expressed positive feedback on ChatGPT's responses to non-clinical questions related to the daily lives of individuals with dementia [307]. This suggests that LLMs can support caregivers by providing accessible and practical information to manage everyday challenges. However, concerns remain about the depth and accuracy of medical information provided by LLMs. Studies comparing ChatGPT with traditional search engines have found limitations in the quality of responses, describing them as accurate but lacking in comprehensiveness [308]. These findings suggest that while LLMs can address basic queries, their applicability in complex medical contexts requires further refinement. One solution to these limitations is fine-tuning LLMs using domain-specific data. For example, models trained in medical journals and textbooks have demonstrated improved performance in handling specialized medical queries [309]. In Alzheimer's research, GPT-4-based tools have been developed to autonomously collect, process, and analyze health information, illustrating how customization can enhance the relevance and precision of information retrieval in specific medical domains [15].

4.2 | Cancer treated by radiation therapy

LLMs have emerged as powerful tools in cancer research, offering innovative solutions for diagnosis, treatment planning, and biological insights. By processing vast datasets of scientific literature, clinical trial results, and genomic information, LLMs can facilitate the identification of novel biomarkers and treatment strategies. LLM-driven multimodal approaches have also enhanced target volume contouring in radiation oncology, integrating imaging data with clinical notes for improved precision [310, 311]. In radiobiology, these models contribute to understanding the complex interplay between radiation and cellular processes, informing the development of personalized treatment regimens [312]. Recent studies also explore the application of LLMs across chemotherapy, surgery, radiotherapy, and immunotherapy, demonstrating their versatility and potential in advancing oncology research.

Multimodal large language models that integrate imaging analysis with NLP have shown promising results in automated organ-at-risk (OAR) and target volume delineation, achieving expert-level performance [313]. These models can process multiple imaging modalities—CT, MRI, and PET—simultaneously while incorporating clinical notes and radiology reports to improve contour accuracy. Additionally, LLMs are being utilized for dose prediction [314], where they have the potential to suggest optimal dose distributions for patients. Recent studies have explored their application in adaptive radiotherapy, where LLMs show potential in processing daily imaging data to recommend plan adaptations based on anatomical changes. Integrating LLMs with knowledge-based planning systems has also enhanced the quality of treatment plans by leveraging insights from large databases of previously treated cases. Furthermore, LLMs demonstrate potential in predicting treatment outcomes and toxicity risks by analyzing patient-specific factors, enabling more personalized treatment approaches.

In clinical practice, LLMs are proving useful in automating routine tasks and supporting complex decision-making [48]. Tools such as ChatGPT have been piloted for generating comprehensive patient case reports, improving the efficiency of clinical documentation. Furthermore, LLMs have shown promise in extracting discrete data elements from clinical notes, aiding in the creation of robust cancer databases. They were evaluated for supporting personalized oncology by recommending clinical trials for head and neck cancer and offering decision support for treatment planning. However, these applications require rigorous validation to ensure the accuracy and reliability of their outputs. In education, LLMs are transforming how knowledge is disseminated and acquired in oncology. Educational chatbots tailored to radiation oncology can simulate patient interactions, helping trainees refine their communication skills [315]. Additionally, LLMs assist in evaluating radiotherapy plans and providing structured feedback, as demonstrated by recent studies. These models foster a more interactive and adaptive learning environment, enabling personalized educational experiences for medical physicists, oncologists, and other healthcare professionals [316]. Despite challenges such as ensuring content accuracy and avoiding the propagation of biases, the integration of LLMs into educational frameworks holds the potential to enhance competency and foster innovation in cancer care.

4.3 | Infectious diseases

4.3.1 | Disease prediction and vaccine efficacy analysis

LLMs such as GPT-3 and GPT-4, have emerged as powerful tools in disease prediction and vaccine efficacy analysis. By processing vast datasets, including biomedical records and epidemiological trends, LLMs can model the spread of infectious diseases, predict vaccination outcomes, and assist in assessing vaccine effectiveness. For example, neural networks combined with logistic regression have been applied to predict influenza vaccination outcomes, achieving significant accuracy based on demographic and clinical data [317]. In the context of pediatric respiratory diseases, ChatGPT has been used to generate insights and recommendations for reducing severe cases post-COVID-19, highlighting the adaptability of LLMs in addressing real-world healthcare issues [318]. Additionally, machine learning algorithms based on clinical features have been validated for predicting influenza infection in patients with influenza-like illness, illustrating the role of LLMs in early diagnosis and targeted intervention [319]. LLMs are also instrumental in identifying immune biomarkers that predict vaccine responsiveness, as seen in studies exploring apoptosis and other immune markers to assess influenza vaccine efficacy [320]. Furthermore, LLMs have been applied to the extraction and analysis of post-marketing adverse events from the Vaccine Adverse Event Reporting System (VAERS), providing valuable insights into vaccine safety and public health implications [321]. The use of machine learning for seasonal antigenic prediction, particularly for influenza A H3N2, demonstrates LLMs' potential in tracking viral evolution and optimizing vaccine design to address emerging strains [322]. As LLM technology continues to advance, its application in disease prediction and vaccine efficacy is expected to become increasingly essential in public health management and disease prevention strategies.

4.3.2 | Vaccine adherence and risk prediction

Machine learning and feature selection techniques, facilitated by LLMs, are essential in analyzing vaccine adherence patterns and identifying factors influencing vaccination rates. These methods allow researchers to process large, complex datasets, uncovering demographic and health-related variables that impact vaccine adherence and risk prediction. For example, machine learning models have been applied to assess low adherence to influenza vaccination among adults with cardiovascular disease, offering insights into the unique barriers to vaccination faced by high-risk groups [323]. Real-time data from online self-reports, such as social media posts, have also been used to track influenza vaccine uptake, providing valuable insights into public sentiment and adherence trends [324]. Furthermore, sociodemographic predictors of vaccine acceptance, especially during the COVID-19 pandemic, have been studied extensively. For instance, machine learning has been used to explore the influence of variables such as education level, income, and geographic location on vaccine hesitancy across various populations [325]. In addition, validated scales such as the parental attitude about childhood vaccination scale have been enhanced with feature selection techniques, refining our understanding of factors associated with vaccine acceptance and hesitancy [326]. Other studies emphasize the broader implications of vaccine hesitancy by analyzing attitudes toward COVID-19 vaccinations across continents, highlighting the variability in hesitancy due to cultural and regional factors [327]. Lastly, comparative studies on flu vaccine uptake pre- and post-COVID-19 leverage machine learning to identify shifts in adherence patterns and factors that predict vaccination behavior over time [328]. Together, these advancements in machine learning and feature selection provide a comprehensive understanding of vaccine adherence, informing targeted public health strategies to improve vaccination rates.

4.3.3 | Biomarker analysis and antigen prediction

LLMs and machine learning approaches are increasingly being applied to analyze biomarkers and predict antigenic variations, which are essential for understanding immune responses and optimizing vaccine design. In biomarker analysis, studies have leveraged LLMs to investigate genetic relationships and autoimmune markers, helping to elucidate the factors that influence vaccination outcomes and susceptibility to infectious diseases [329]. For example, models have been employed to identify key susceptibility hubs within biological networks, offering insights into factors that contribute to immune response variability [330].

Additionally, antigenic prediction plays a crucial role in designing effective influenza vaccines, especially for rapidly evolving strains. Statistical analyses of antigenic similarity, such as those conducted for influenza A (H3N2), highlight the potential of machine learning models in mapping antigenic drift and optimizing strain selection for seasonal vaccines [331]. Moreover, cellular correlates of protection identified through human influenza virus challenges have advanced our understanding of immune responses to oral vaccines, demonstrating the applicability of machine learning models in immune signature identification [332]. Blood inflammatory biomarkers have also been analyzed to differentiate COVID-19 from influenza cases, showcasing the predictive power of LLMs in clinical biomarker differentiation [323]. Seasonal antigenic prediction, particularly for influenza A H3N2, has benefited from machine learning approaches that help forecast viral evolution, supporting timely vaccine updates [322]. Finally, phylogenetic analyses have identified optimal influenza virus candidates for seasonal vaccines, underscoring the significance of LLMs in guiding vaccine development against anticipated strains [333].

4.3.4 | Vaccine recommendation and immune response

LLMs are increasingly leveraged in vaccine recommendation and immune response studies, especially in analyzing antigenicity and optimizing vaccine strain selection. For instance, the MAIVeSS platform utilizes LLMs to streamline the selection of high-yield, antigenically matched viruses for seasonal influenza vaccines, a critical step in addressing annual viral mutations [334]. In populations with specific health conditions, such as human immunodeficiency virus (HIV), LLMs have been applied to predict the immunogenicity of trivalent inactivated influenza vaccines, revealing key biomarkers and immune signatures that inform personalized vaccination strategies [335].

Antigenicity prediction models have also employed convolutional neural networks to optimize vaccine recommendations for influenza virus A (H3N2), facilitating the identification of effective vaccine strains through detailed computational modeling [336]. Furthermore, temporal topic models generated from clinical text data allow for a more nuanced understanding of immune responses over time, especially in relation to patient health history and demographic factors, enhancing the precision of vaccine recommendations [337]. Finally, studies on COVID-19 vaccine hesitancy among populations already immunized for influenza underscore the relevance of LLMs in analyzing and addressing hesitancy factors, which is vital for improving adherence to vaccination programs [338]. Together, these applications illustrate the potential of LLM-based approaches in advancing vaccine recommendation processes and tailoring immune response strategies.

4.3.5 | Sentiment analysis and public attitude research on social media

LLM techniques are widely used in sentiment analysis to assess public attitudes toward vaccines, particularly through social media data. This approach provides insights into public sentiment trends and identifies factors contributing to vaccine hesitancy or acceptance. For instance, social media analysis of public messaging around influenza vaccination from 2017 to 2023 has shown how sentiment fluctuates in response to vaccine news, policy changes, and health crises, offering a longitudinal view of public perception [339]. Similarly, negative sentiments related to influenza vaccines, analyzed from over 260,000 Twitter posts, highlight recurring concerns and misconceptions that can be addressed through targeted public health messaging [340].

Beyond social media, predictive models using smartwatch and smartphone data can monitor side effects and public reactions post-vaccination, enhancing our understanding of vaccine safety perceptions [341]. The FDA's Biologics Effectiveness and Safety Initiative also uses NLP to process unstructured data, identifying adverse events associated with vaccines and contributing to more accurate public health responses [342]. Additionally, integrating immune cell population data and gene expression with CpG methylation patterns offers insights into immune responses that can correlate with public attitudes, informing data-driven interventions [343]. These findings underscore the utility of LLMs in sentiment analysis, enabling public health authorities to monitor and respond to vaccine-related concerns effectively.

4.3.6 | Epidemiology and public health data analysis

Machine learning and large datasets have profoundly impacted epidemiology and public health, enabling the analysis of disease patterns, risk factors, and vaccination responses. Studies integrating socioeconomic, health, and safety data have examined how these factors affect COVID-19 spread, offering insights into the influence of demographics such as income and healthcare access on infection rates [344]. Projects such as the Human Vaccines Project also leverage large datasets to map immune responses across populations, enhancing our understanding of vaccine design and immunology [345].

The use of wearable sensors in epidemiological studies, as demonstrated in the WE SENSE protocol, facilitates early detection of viral infections by analyzing real-time health metrics, thus supporting timely public health interventions [346]. Pneumonia research, such as the work by the CAPNETZ study group, highlights unmet needs in understanding disease mechanisms, emphasizing the need for targeted data collection and analysis in developing effective treatment and intervention strategies [347]. Additionally, sociodemographic studies on COVID-19 vaccine acceptance reveal how age, gender, and education level impact vaccine uptake, providing crucial insights for public health policy [348]. These applications underscore the essential role of data-driven approaches in epidemiology and public health to improve disease prevention and health policy.

5 | DRUG DISCOVERY AND DEVELOPMENT

5.1 | Drug target identification

Drug discovery is a resource-intensive and time-consuming process, often spanning 7–20 years from initial development to market approval [349, 350]. Central to this process is DTI identification, which involves pinpointing molecules implicated in disease mechanisms. Traditional methods, including genomics, proteomics, RNAi, and molecular docking, have been instrumental but face limitations in cost, scalability, and adaptability to complex biological systems (Figure 2).

Recent advancements in computational techniques, such as machine learning [351–353], knowledge graph-based methods [354, 355], and molecular docking simulations, driven by the rapid growth of large-scale biomedical datasets [356–358], have significantly advanced DTI prediction. Beyond these methods, recent breakthroughs in LLMs and bioinformatics-specific language models represent a paradigm shift, enabling the integration and analysis of vast, heterogeneous datasets—including molecular data, biological networks, and scientific literature—while storing drug-related background knowledge through extensive pretraining [359–363]. This section provides an overview of LLM-based approaches for DTI prediction, categorized based on the type of data they utilize sequence data, structural data, and relationship data, with the latter primarily derived from knowledge graphs.

Sequence data, including amino acid sequences for proteins and SMILES representations for drugs, plays a central role in single-modal methods for DTI prediction. Pretrained language models (PLMs), such as PharmBERT [364], BioBERT [174], and ProteinBERT [235], have been widely utilized to extract meaningful representations from such data, enabling efficient and accurate predictions. For instance, DTI-LM [365] addresses the cold-start problem by utilizing PLMs to predict DTIs based solely on molecular and protein sequences, enabling accurate predictions for novel drugs and uncharacterized targets. Similarly, ConPLex [366] generates co-embeddings of drugs and target proteins, achieving broad generalization to unseen proteins and over 10× faster inference compared to traditional sequence-based methods, making it ideal for tasks such as drug repurposing and high-throughput screening. Yang et al. [367] further enhance DTI prediction by introducing high-frequency amino acid subsequence embedding and transfer learning, capturing functional interaction units and shared features across large datasets. Additionally, TransDTI [368] employs transformer-based language models to classify drug-target interactions into active, inactive, or intermediate categories, offering competitive performance. Despite their advantages, single-modal methods are limited by their reliance on sequence data alone, making it challenging to capture interactions involving spatial, structural, or contextual dependencies.

To address the limitations of single-modal approaches, multimodal frameworks integrate diverse data types—such as molecular graphs, protein sequences, and structural data—offering a more comprehensive understanding of DTIs. DrugLAMP [369] exemplifies this integration, utilizing pocket-guided co-attention and paired multi-modal attention to fuse molecular graphs with sequence data, achieving nuanced molecular interaction predictions. PGraphDTA [370] incorporates 3D contact maps alongside protein sequences, outperforming sequence-only methods when structural data is available. Beyond predictive accuracy, multimodal frameworks such as CGPDTA [371] enhance interpretability by integrating interaction networks, providing insights into biological mechanisms. DrugChat [372] combines prompt-based learning with sequence data and textual inputs. Pretrained on three datasets, it predicts indications, mechanisms of action, and pharmacodynamics while dynamically generating textual outputs in response to user prompts. This eliminates the need for retraining and enables flexible, interactive exploration of drug mechanisms. Similarly, DrugReAlign [373] employs a multi-source prompting approach that integrates diverse and reliable data inputs to integrate textual and structural data, enhancing drug repurposing efforts.

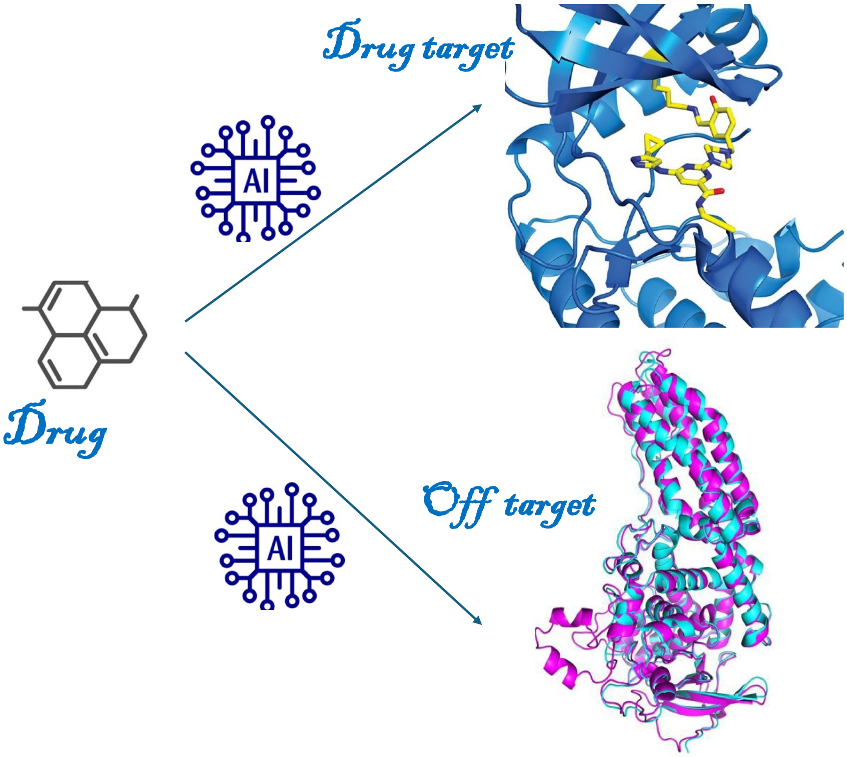

FIGURE 2 Schematic diagram of drug target.

Beyond structural data, KG-based models leverage semantic relationships, such as shared pathways, biological processes, and functional annotations, along with diverse data sources to achieve competitive performance in DTI predictions. Y-Mol [373] enhances biomedical reasoning by integrating multiscale biomedical knowledge and using LLaMA2 as its base LLM. It learns from publications, knowledge graphs, and synthetic data, enriched by three types of drug-oriented prompts: description-based, semantic-based, and TB, enabling robust drug interaction analysis. Similarly, the multi-agent framework DrugAgent [374] advances drug repurposing by combining AI-driven DTI models, knowledge graph extraction from databases (e.g., DrugBank, CTD [375]), and literature-based validation. This framework integrates diverse data sources to streamline repurposing candidate identification, enhancing efficiency, interpretability, and cost-effectiveness. Together, these models boost predictive power while fostering collaboration and refinement.

5.2 | Molecular docking and drug design

The advanced reasoning capabilities of LLMs have enabled their application in biological and medical fields, demonstrating significant potential to accelerate drug discovery and screening processes [17, 73]. Built upon the transformer architecture from NLP, biology-focused language models have emerged as powerful tools to support both sequence-based and structure-based drug design (SBDD) [376–378]. By utilizing their strengths in TS and contextual understanding, these models can integrate information from diverse sources, such as scientific literature, patent databases, and specialized datasets, to provide comprehensive analyses and insights into protein sequences, structures, binding pockets, and interaction sites [379]. Moreover, protein language models and other transformer-based models are being applied to exploit unknown structural information in SBDD [377, 378].

Molecular docking, a pivotal component of SBDD, necessitates three-dimensional protein structures and precise binding site information to calculate binding affinities during silico virtual screening [380]. LLMs have shown potential to enhance various aspects of molecular docking, including docking input file generation, binding site prediction, and protein structure prediction [377, 378, 381]. AutoDock is a widely adopted software for molecular docking [382]. For high-throughput drug screening, it is necessary to generate docking commands in text file format and execute them in the terminal. Sharma et al. demonstrated the capability of ChatGPT to generate AutoDock input files and basic molecular docking scripts [381]. Another notable example is DrugChat, a ChatGPT-like LLM for drug molecule graphs developed by Liang et al. With the input of compound molecule graphs and appropriate prompts, DrugChat is able to generate insightful responses [383].

Ligand binding site identification and prediction are essential for drug design. Due to the limited availability of experimentally determined protein crystal structures and incomplete protein structural knowledge, ligand binding site identification can be tough. Zhang and Xie addressed this limitation through LaMPSite, an algorithm powered by EMS-2 protein language model, which only requires protein sequences and ligand molecular graphs as inputs without any protein structural information [377]. This approach achieved comparable performance to those methods requiring 3D protein structures in benchmark evaluations. Regarding deficiency of reliable protein structure, protein language models have been applied for protein structure prediction as well. For example, Fang et al. introduced HelixFold-Single, a multiple-sequence-alignment-free protein structure predictor [378]. Unlike AlphaFold2, which enhances prediction accuracy by relying on MSAs of homologous proteins, HelixFold-Single adopts a more efficient approach. It leverages large-scale protein language model training on the primary structures of proteins while integrating key components from AlphaFold2 for protein geometry.

Recent advancements in protein-ligand binding prediction methods have further enhanced screening efficiency and accuracy. Shen et al. developed RTMScore, which integrated graph transformer to extract structural features of protein and molecule, using 3D residue graphs of protein and 2D molecular graphs as inputs for protein-ligand binding pose prediction [384]. RTMScore outperformed many SOTA docking software including Autodock Vina [385], DeepBSP [386], and DeepDock [387] in performing virtual screening tasks. Another notable development is ConPlex, a sequence-based DTI prediction method introduced by Singh et al. [366]. By employing representations generated from pre-trained protein language models as the inputs, ConPlex benefits from a larger corpus of single protein sequences and alleviates the problem of limited DTI training data. Additionally, contrastive learning was adopted to address the fine-grained issues by employing contrastive coembedding, which is able to co-locate the proteins and the targets in a shared latent space. Thus, a high specificity can be achieved by separating the true interacting patterns and decoys. According to contrastive training results, the effective size between true and decoy scores was largely increased.

Through automated data extraction and normalization, LLMs can greatly improve the efficiency and accuracy of drug property predictions. With absorption, distribution, metabolism, excretion, and toxicity (ADMET) analysis, LLMs can also help distinguish the compounds possessing favorable profiles from those showing adverse characters and allow developing the most promising drug candidates during the pipeline process. For instance, PharmaBench achieves this through its multi-intelligence system, whose core function is to extract ADMET-related data from multiple public databases using LLMs [388]. Beyond ADMET analysis, LLMs such as ChatGPT have expanded their capabilities to predict and analyze other features of drugs, including pharmacodynamics and pharmacokinetics, thus providing a comprehensive evaluation of potential drug candidates [379]. LLMs powerfully accelerate the drug development pipeline by fastening data analysis, enhancing prediction accuracy, and offering all-rounded drug property evaluation, which in turn reduces both the time and resources needed for drug discovery and improves the chances of coming up with a successful drug candidate.

6 | IMMUNOLOGY AND VACCINE DEVELOPMENT

LLMs, including GPT-based architectures, have transformed the field of immunology and vaccine development by enabling advanced analyses of large, complex datasets. These models, combined with machine learning, NLP, and feature selection techniques, facilitate the identification of immune biomarkers, prediction of vaccine efficacy, understanding of vaccine hesitancy, and real-time monitoring of adverse events. This review synthesizes recent research highlighting the critical role of LLMs in advancing vaccine science, with a focus on immune response analysis, vaccine development, efficacy prediction, safety, and public attitudes.

6.1 | Immune response analysis and biomarker research

Analyzing immune responses and identifying biomarkers are critical for understanding the efficacy and mechanisms of vaccines. LLMs, integrated with advanced computational techniques, play a key role in processing and interpreting complex datasets to uncover immune signatures and their correlation with vaccination outcomes. For example, LLMs can efficiently analyze high-dimensional datasets, such as the FluPRINT dataset, which provides a multidimensional analysis of the immune system's imprint following influenza vaccination, revealing variability in immune responses across individuals [389]. By leveraging LLMs, researchers can extract patterns and relationships from immune cell populations, mRNA sequencing, and CpG methylation data, leading to more accurate predictions of humoral immunity and highlighting the impact of gene expression and epigenetic modifications on vaccine-induced immunity [390].

Automated systems such as SIMON utilize machine learning, augmented by LLMs for text-based data extraction and interpretation, to reveal immune signatures that predict vaccine responsiveness, providing deeper insights into immune mechanisms [391]. Furthermore, LLMs facilitate the integration of multilevel models that incorporate gene expression interaction networks to predict antibody responses to vaccines, enabling the precise identification of immune predictors [392]. For biomarker analysis, LLMs contribute to identifying apoptosis and inflammatory responses through their ability to process vast quantities of biological literature and experimental data, as seen in studies linking immune biomarkers with influenza vaccine responsiveness [390]. They also assist in differentiating immune responses to COVID-19 and influenza infections by analyzing blood inflammatory biomarkers and clinical data at scale [393].

Additionally, human influenza virus challenge models, supported by LLM-driven analysis of experimental outcomes, have identified cellular correlates of protection, advancing our understanding of immune responses to oral vaccines [332]. LLMs streamline the analysis of complex immune response datasets, ensuring faster identification of key findings and improving collaboration across interdisciplinary research teams.

6.2 | Vaccine development and recommendation models

The development and optimization of vaccines rely on computational models to predict vaccine efficacy, identify suitable strains, and recommend antigenically matched candidates. LLMs have become invaluable tools in this domain by enhancing the ability to process and analyze vast datasets, extract patterns from biomedical literature, and improve antigenic prediction models. Neural networks and logistic regression have traditionally been applied to predict influenza vaccination outcomes, providing robust frameworks for assessing vaccine effectiveness based on demographic and clinical data [389]. With the integration of LLMs, these predictive models can be further refined by incorporating insights derived from textual datasets, such as clinical notes, trial reports, and patient feedback.

In silico approaches, combined with LLM-based text mining, enable the analysis of autoimmune diseases and their genetic relationships to vaccination. LLMs can extract relevant patterns across large corpora of genomic and immunological studies, offering deeper insights into immune response mechanisms and potential cross-reactivity among populations [394].

Platforms such as MAIVeSS streamline the selection of antigenically matched, high-yield viruses for seasonal influenza vaccines by leveraging LLMs to analyze historical viral sequences, antigenic relationships, and experimental outcomes [334]. Additionally, seasonal antigenic prediction models utilize machine learning algorithms integrated with LLMs to analyze influenza A (H3N2) evolution and forecast emerging strains, improving the accuracy and efficiency of vaccine formulation [322].

Phylogenetic analyses are also augmented through LLM capabilities, which automate literature reviews and contextualize genetic relationships to identify influenza virus candidates for seasonal vaccines. This ensures antigenic compatibility, reduces manual analysis time, and maximizes immunogenic coverage [333]. By incorporating LLMs, researchers can process and synthesize global influenza surveillance data, generating actionable insights to address the challenge of rapidly evolving pathogens.

6.3 | Vaccine efficacy prediction and immunogenicity studies

Accurately predicting vaccine efficacy and assessing immunogenicity are critical for improving vaccination strategies and understanding immune responses. LLMs play a pivotal role in processing vast datasets to extract critical insights, identify risk factors, and predict vaccine efficacy. LLMs are increasingly used to synthesize clinical, epidemiological, and behavioral data, which are key to identifying populations with low adherence to vaccination programs. For example, models analyzing high-risk groups, such as individuals with cardiovascular disease, have integrated LLM-driven data extraction from clinical records to uncover demographic and behavioral predictors of vaccine uptake [395].

In real-time monitoring, LLMs enhance the analysis of self-reported data to estimate vaccine coverage and adherence. By processing text-based survey responses and digital health data, LLMs enable precise insights into population-wide vaccine uptake and the factors influencing these trends [392].

For immunogenicity studies, LLMs are employed to mine complex biological and clinical datasets, improving predictions of vaccine immune responses in targeted populations. For example, LLM-augmented artificial intelligence models have been used to predict immunogenicity in pediatric studies, such as for trivalent inactivated influenza vaccines in HIV-infected children, facilitating personalized vaccination strategies [396]. Clinical feature-based models further benefit from LLMs' ability to extract structured and unstructured data from clinical notes, improving predictions of infection risks in individuals' post-vaccination [319].

In biomarker-based analyses, LLMs assist in synthesizing large-scale experimental and clinical literature to identify apoptosis markers and inflammatory biomarkers associated with vaccine responsiveness. This enables a better understanding of immune responses and facilitates personalized immunization approaches [397]. Post-marketing vaccine safety surveillance systems, such as the VAERS dataset, have leveraged LLMs to extract, classify, and analyze adverse event reports. By automating the processing of unstructured clinical narratives, LLMs enhance the detection of adverse events and improve vaccine safety assessments [325].

Comparative studies examining influenza vaccine uptake pre- and post-COVID-19 also benefit from LLMs' ability to analyze large textual datasets, such as survey responses and social media discussions. These models provide actionable insights into behavioral shifts and critical predictors of vaccine adherence, contributing to data-driven vaccination strategies [328].

6.4 | Vaccine hesitancy and public attitude analysis

Vaccine hesitancy remains a significant challenge to achieving widespread immunization, and LLMs have proven instrumental in uncovering the underlying causes, trends, and predictors of public attitudes toward vaccination. LLMs, combined with machine learning and NLP, enable the analysis of large-scale textual data, including social media, survey responses, and clinical reports, providing insights into public perceptions and vaccine acceptance patterns.

LLMs have been employed to process unstructured data for real-time monitoring of vaccine-related discussions, identifying concerns around side effects and safety perceptions. For example, predictive models using smartwatch and smartphone data, enhanced by LLM-driven text analysis, have been used to detect and predict the severity of side effects following vaccination, improving the understanding of public concerns regarding vaccine safety [341]. LLMs have further facilitated automated detection of vaccine-related messaging and adverse event reporting, as demonstrated by initiatives such as the FDA Biologics Effectiveness and Safety Initiative [342]. These models analyze clinical notes and text-based reports at scale, streamlining post-vaccination safety monitoring.

Parental attitudes toward childhood vaccination have been analyzed using validated scales, with LLMs efficiently extracting themes and patterns from caregiver responses. These analyses highlight key concerns, such as vaccine safety and efficacy, and inform targeted education strategies [326]. LLMs are also applied to sociodemographic studies, enabling the identification of key predictors of vaccine acceptance, including education level, income, and geographic location. By synthesizing national-scale survey data, LLMs provide a foundation for interventions aimed at addressing vaccine hesitancy in specific demographic groups [389].

Sentiment analysis of social media platforms, such as Twitter, has been revolutionized by LLMs such as GPT-based architecture. These models analyze vaccine-related discourse, identifying trends in vaccine hesitancy and negative attitudes toward vaccination programs. For example, LLMs have revealed hesitancy trends related to influenza vaccination and highlighted shifts in public sentiment in response to public health campaigns and policy changes [339, 340]. Longitudinal studies powered by LLMs demonstrate how public messaging evolved over multiple years, providing actionable insights for optimizing communication strategies and combating misinformation. Additionally, comparative studies across continents emphasize cultural and regional variations in vaccine attitudes, which LLMs can analyze to tailor communication strategies to local contexts [327].

Vaccine hesitancy studies among specific groups, such as Canadians immunized for influenza, benefit from LLMs' ability to process large-scale survey responses and extract nuanced concerns [338]. These insights underscore the complexity of public attitudes and the importance of sustained public education.

6.5 | Vaccine safety and adverse event detection

Ensuring vaccine safety and monitoring adverse events following immunization are critical components of immunization programs. LLMs play an increasingly vital role in enhancing vaccine safety surveillance by automating the detection, classification, and analysis of adverse events at scale.

Predictive models leveraging smartwatch and smartphone data, combined with LLM-powered text analysis, enable real-time monitoring of vaccine-related side effects. By processing unstructured patient-reported outcomes and wearable device data, LLMs help identify patterns in the severity of side effects following COVID-19 and influenza vaccinations, facilitating timely interventions and improving patient outcomes [341].

The FDA's Biologics Effectiveness and Safety Initiative has utilized NLP techniques powered by LLMs to analyze unstructured clinical data, such as physician notes and medical records, for detecting vaccine-related adverse events. LLMs significantly enhance the ability to process and interpret large-scale textual datasets, streamlining the identification of safety signals and improving the efficiency of post-marketing surveillance systems [398]. These automated systems reduce manual effort, accelerate safety signal detection, and enable regulators to respond quickly to emerging concerns.

Deep learning approaches applied to the VAERS have also been strengthened by the integration of LLMs. By extracting and categorizing adverse event reports from free-text submissions, LLMs improve the accuracy and granularity of vaccine safety assessments. For instance, LLMs can identify subtle patterns and correlations within adverse event reports, allowing researchers to generate valuable insights into vaccine safety profiles, detect rare adverse events, and support regulatory decisions [399].

Moreover, LLMs facilitate cross-referencing of adverse event data with other sources, such as scientific literature, clinical trial reports, and patient feedback, providing a comprehensive view of vaccine safety. By automating this process, LLMs improve the robustness of post-marketing surveillance systems and enhance public confidence in vaccination programs.

Together, these studies underscore the transformative role of LLMs in vaccine safety monitoring and adverse event detection. By efficiently processing and analyzing large-scale unstructured data, LLMs enable faster, more accurate identification of adverse events, ensuring vaccine safety and maintaining public trust in immunization efforts.

6.6 | Vaccine-related social and health data analysis

The analysis of social and health data plays a critical role in understanding vaccine uptake, disease spread, and public health outcomes. LLMs have become essential tools for processing and interpreting large-scale social, health, and demographic datasets, enabling researchers to identify patterns and design targeted interventions for improving vaccination strategies.

LLMs are particularly effective in synthesizing diverse data sources, including EHRs, socioeconomic surveys, and real-time reports. For instance, studies addressing unmet needs in pneumonia research benefit from LLMs' ability to integrate textual clinical data with structured epidemiological datasets, facilitating a comprehensive understanding of disease burden, treatment gaps, and prevention strategies [347]. Similarly, LLMs assist in analysing socioeconomic, health, and safety data to explain the spread of diseases such as COVID-19. By processing large datasets across regions, LLMs highlight key factors, such as healthcare access, education, and income, that impact infection rates and vaccine coverage [400].

The Human Vaccines Project, which focuses on leveraging immunological and epidemiological data to improve vaccination strategies, has incorporated LLMs to process vast volumes of immunology literature, trial reports, and population health data. LLMs enable the identification of critical trends and insights that advance the understanding of human immune responses to vaccines, accelerating the development of targeted immunization programs [390].

Sociodemographic predictors of vaccine acceptance, such as age, education, and geographic location, have been extensively studied with the support of LLMs. These models efficiently process large-scale national surveys, extracting patterns and correlations that inform targeted interventions to address vaccine hesitancy and acceptance across diverse populations [389].

Real-time health data collected through wearable sensors, as demonstrated in the WE SENSE protocol, have also been enhanced by LLMs' ability to integrate sensor outputs with epidemiological trends. LLMs process this real-time data alongside other health records to detect early warning signs of viral infections and identify potential outbreaks, highlighting their role in improving public health preparedness and surveillance [346].

These studies collectively demonstrate the transformative role of LLMs in integrating social, health, and demographic data for vaccine-related research. By efficiently processing and analyzing vast, heterogeneous datasets, LLMs offer valuable insights that shape public health policies, improve vaccination strategies, and enhance disease preparedness efforts.

7 | DISCUSSION AND FUTURE DIRECTIONS

Although LLMs have achieved remarkable success in bioinformatics, they still face numerous challenges. The performance of LLMs in bioinformatics heavily relies on the quality of training data, yet available datasets such as genomic or proteomic sequences often contain noise and biases. This issue leads to inaccurate predictions and limited generalizability. Additionally, the limited availability of labeled biological data further hinders the adaptability of LLMs to diverse bioinformatics tasks. Computational cost and scalability present another significant challenge. LLMs are resource-intensive, requiring substantial computational power and memory for training and inference, which becomes particularly problematic when analyzing ultra-long sequences such as genomic regions spanning thousands of base pairs. Transformer-based architectures, despite their advancements, still struggle with scaling efficiently for such long sequences due to inherent memory constraints.

Generalizability and interpretability also remain critical concerns. While LLMs excel at specific tasks, their ability to generalize across unseen datasets or tasks is often inadequate. Moreover, the lack of interpretability in model output makes it difficult for researchers to understand the underlying biological mechanisms, which are essential for result validation. Ethical and privacy concerns further complicate the application of LLMs, particularly in sensitive areas such as personalized medicine. The use of patient data in training models raises significant ethical questions and potential privacy risks, limiting widespread adoption.

Despite these challenges, the future of LLMs in bioinformatics presents exciting opportunities. Efforts are likely to focus on developing lightweight and efficient architectures, such as LoRA and QLoRA, to mitigate computational and memory requirements. Innovations in transformer variants and hybrid architecture are expected to overcome scalability challenges, enabling more effective analysis of long-sequence bioinformatics tasks. Integrating diverse biological data types, including DNA, RNA, protein sequences, epigenetic, and transcriptomic data, will enhance LLMs' capability to generate comprehensive biological insights. Improved interpretability will also become a priority, with advancements aimed at visualizing attention mechanisms and uncovering the biological basis behind predictions.

Applications in personalized medicine highlight the transformative potential of LLMs. For example, they can revolutionize precision medicine by tailoring treatments to individual patients, predicting drug efficacy, or identifying possible side effects based on genomic data. Addressing data scarcity through open data initiatives and interdisciplinary collaborations will further accelerate progress, enabling broader applications of LLMs in bioinformatics. Additionally, as Transformer models reach maturity, exploration of alternative architectures may drive innovation beyond their current limitations, ensuring continuous advancement in the field. These trends underscore the dynamic evolution of LLMs in bioinformatics, presenting opportunities for groundbreaking developments while emphasizing the need to address existing limitations.

The integration of multimodal biomedical data presents another promising direction for future research. Sequence-to-sequence models, which have demonstrated remarkable success in NLP, offer a promising technical approach for fusing diverse biomedical data types. These models can potentially bridge the gap between different modalities—including medical imaging, clinical texts, temporal data (such as EHRs and vital signs), and various forms of biological sequence data (DNA, RNA, and proteins). For instance, sequence-to-sequence architectures could be adapted to translate between modalities [17, 401], such as converting radiological images to diagnostic text descriptions while incorporating relevant genomic information. This multimodal fusion could enable more comprehensive disease diagnosis and treatment planning by leveraging complementary information from different data sources. Furthermore, innovative attention mechanisms and cross-modal transformers could help capture complex relationships between different data types, leading to more robust and interpretable models. The challenge lies in developing architectures that can effectively handle the inherent heterogeneity of these data types while maintaining computational efficiency and biological interpretability.

8 | CONCLUSION

This comprehensive survey has explored the transformative impact of LLMs in bioinformatics, spanning applications in genomics, proteomics, drug discovery, and clinical medicine. Our review has highlighted the successful adaptation of transformer architectures for biological sequences, the emergence of specialized biomedical LLMs, and the promising integration of multiple data modalities. These advances have enabled significant progress in protein structure prediction, DTI analysis, and disease diagnosis.

Despite notable achievements, challenges persist in data quality, computational scalability, model interpretability, and ethical considerations regarding patient privacy. These challenges present opportunities for future research, particularly in developing efficient architectures, improving multimodal data integration, and ensuring model interpretability. The convergence of LLMs with emerging biotechnologies promises to accelerate discovery in bioinformatics, potentially leading to more precise and personalized medical interventions.

AUTHOR CONTRIBUTIONS

Wei Ruan: Writing—original draft; writing—review and editing. **Yanjun Lyu:** Writing—original draft; writing—review and editing. **Jing Zhang:** Writing—original draft; writing—review and editing. **Jiazhang Cai:** Writing—original draft. **Peng Shu:** Writing—original draft. **Yang Ge:** Writing—original draft. **Yao Lu:** Writing—original draft. **Shang Gao:** Writing—original draft. **Yue Wang:** Writing—original draft. **Peilong Wang:** Writing—original draft. **Lin Zhao:** Writing—original draft. **Tao Wang:** Writing—original draft. **Yufang Liu:** Writing—original draft. **Luyang Fang:** Writing—original draft. **Ziyu Liu:** Writing—original draft. **Zhengliang Liu:** Writing—original draft. **Yiwei Li:** Writing—original draft. **Zihao Wu:** Writing—original draft. **Junhao Chen:** Writing—original draft. **Hanqi Jiang:** Writing—original draft. **Yi Pan:** Writing—original draft. **Zhenyuan Yang:** Writing—original draft. **Jingyuan Chen:** Writing—original draft. **Shizhe Liang:** Writing—original draft. **Wei Zhang:** Writing—original draft. **Terry Ma:** Writing—original draft. **Yuan Dou:** Writing—original draft.

Jianli Zhang: Writing—original draft. **Xinyu Gong:** Writing—original draft. **Qi Gan:** Writing—original draft. **Yusong Zou:** Writing—original draft. **Zebang Chen:** Writing—original draft. **Yuanxin Qian:** Writing—original draft. **Shuo Yu:** Writing—original draft. **Jin Lu:** Writing—review and editing. **Kenan Song:** Writing—review and editing. **Xianqiao Wang:** Writing—review and editing. **Andrea Sikora:** Writing—review and editing. **Gang Li:** Writing—review and editing. **Xiang Li:** Writing—review and editing. **Quanzheng Li:** Writing—review and editing. **Yingfeng Wang:** Writing—review and editing. **Lu Zhang:** Writing—review and editing. **Yohannes Abate:** Writing—review and editing. **Lifang He:** Writing—review and editing. **Wenxuan Zhong:** Writing—review and editing. **Rongjie Liu:** Writing—review and editing. **Chao Huang:** Writing—review and editing. **Wei Liu:** Writing—review and editing. **Ye Shen:** Writing—review and editing. **Ping Ma:** Writing—review and editing. **Hongtu Zhu:** Writing—review and editing. **Yajun Yan:** Writing—review and editing. **Dajiang Zhu:** Writing—review and editing. **Tianming Liu:** Writing—review and editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ETHICS STATEMENT

This article does not contain any studies with human or animal materials performed by any of the authors.

REFERENCES

[1] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, Minneapolis, Minnesota, 1; 2019. p. 2.

[2] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.

[3] Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. Large language models in bioinformatics: applications and perspectives. 2024. Preprint at arXiv: 2401.04155v1.

[4] Sarumi OA, Heider D. Large language models and their applications in bioinformatics. Comput Struct Biotechnol J. 2024;23:3498–505.

[5] Tripathi S, Gabriel K, Tripathi PK, Kim E. Large language models reshaping molecular biology and drug development. Chem Biol Drug Des. 2024;103(6):e14568.

[6] Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of chatGPT/GPT-4 research and perspective towards the future of large language models. 2023. Preprint at arXiv: 2304.01852.

[7] Lin Z, Zhang L, Wu Z, Chen Y, Dai H, Yu X, et al. When brain-inspired AI meets AGI. 2023. Preprint at arXiv: 2303.15935.

[8] Zhong T, Liu Z, Pan Y, Zhang Y, Zhou Y, Liang S, et al. Evaluation of openAI o1: opportunities and challenges of AGI. 2024. Preprint at arXiv: 2409.18486.

[9] Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, et al. An iterative optimizing framework for radiology report summarization with chatGPT. IEEE Transactions on Artificial Intelligence. 2024;5(8):4163–75.

[10] Dai H, Liu Z, Liao W, Huang X, Wu Z, Lin Z, et al. ChatAug: leveraging chatGPT for text data augmentation. 2023. Preprint at arXiv: 2302.13007.

[11] Liu Z, Li Y, Peng S, Zhong A, Yang L, Ju C, et al. Radiology-LLaMA2: best-in-class large language model for radiology. 2023. Preprint at arXiv: 2309.06419.

[12] Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y, et al. Differentiating chatGPT-generated and human-written medical texts: quantitative study. JMIR Medical Education. 2023;9(1): e48904.

[13] Liu Z, He X, Liu L, Liu T, Zhai X. Context matters: a strategy to pre-train language model for science education. 2023. Preprint at arXiv: 2301.12031.

[14] Rezayi S, Dai H, Liu Z, Wu Z, Hebbar A, Burns AH, et al. ClinicalRadioBERT: knowledge-infused few shot learning for clinical notes named entity recognition. In: Machine learning in medical imaging: 13th international workshop, MLMI 2022, held in conjunction with MICCAI 2022, Singapore, September 18, 2022, proceedings. Springer; 2022. p. 269–78.

[15] Dai H, Li Y, Liu Z, Lin Z, Wu Z, Song S, et al. AD-autoGPT: an autonomous GPT for Alzheimer's disease infodemiology. 2023. Preprint at arXiv: 2306.10095.

[16] Zhao H, Qian L, Pan Y, Zhong T, Hu J-Y, Yao J, et al. Ophtha-LLaMA2: a large language model for ophthalmology. 2023. Preprint at arXiv: 2312.04906.

[17] Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024;30(11):1–13.

[18] Liu Z, Wang P, Li Y, Holmes J, Peng S, Zhang L, et al. RadOnc-GPT: a large language model for radiation oncology. 2023. Preprint at arXiv: 2309.10160.

[19] Liu Z, Wang P, Li Y, Holmes JM, Peng S, Zhang L, et al. Fine-tuning large language models for radiation oncology, a highly specialized healthcare domain. International Journal of Particle Therapy. 2024;12:100428.

[20] Lyu Y, Wu Z, Zhang L, Zhang J, Li Y, Ruan W, et al. GP-GPT: large language model for gene-phenotype mapping. 2024. Preprint at arXiv: 2409.09825.

[21] Wang J, Jiang H, Liu Y, Ma C, Zhang X, Pan Y, et al. A comprehensive review of multimodal large language models: performance and challenges across different tasks. 2024. Preprint at arXiv: 2408.01319.

[22] Yue H, Sun L, Wang H, Wu S, Zhang Q, Yuan L, et al. Position: trustLLM: Trustworthiness in large language models. In: International conference on machine learning. PMLR; 2024. p. 20166–270.

[23] Liu Z, Zhang L, Wu Z, Yu X, Cao C, Dai H, et al. Surviving chatGPT in healthcare. Frontiers in Radiology. 2024;3:1224682.

[24] Huang Y, Sun L, Wang H, Wu S, Zhang Q, Yuan L, et al. TrustLLM: trustworthiness in large language models. 2024. Preprint at arXiv: 2401.05561.

[25] Yang Z, Liu Z, Zhang J, Lu C, Jiaxin T, Zhong T, et al. Analyzing nobel prize literature with large language models. 2024. Preprint at arXiv: 2410.18142.

[26] Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: methodologies and applications. 2023. Preprint at arXiv: 2304.14670.

[27] Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, et al. Tailoring large language models to radiology: a preliminary approach to LLM adaptation for a highly specialized domain. In: International workshop on machine learning in medical imaging. Springer; 2023. p. 464–73.

[28] Jie T, Hou J, Wu Z, Peng S, Liu Z, Xiang Y, et al. Assessing large language models in mechanical engineering education: a study on mechanics-focused conceptual understanding. 2024. Preprint at arXiv: 2401.12983.

[29] Lee G-G, Shi L, Latif E, Gao Y, Bewersdorf A, Nyaaba M, et al. Multimodality of AI for education: towards artificial general intelligence. 2023. Preprint at arXiv: 2312.06037.

[30] Peng S, Zhao H, Jiang H, Li Y, Xu S, Pan Y, et al. LLMs for coding and robotics education. 2024. Preprint at arXiv: 2402.06116.

[31] Latif E, Mai G, Nyaaba M, Wu X, Liu N, Lu G, et al. Artificial general intelligence (AGI) for education. 2023. Preprint at arXiv: 2304.12479, 1.

[32] Wang J, Wu Z, Li Y, Jiang H, Peng S, Shi E, et al. Large language models for robotics: opportunities, challenges, and perspectives. 2024. Preprint at arXiv: 2401.04334.

[33] Liu Y, He H, Han T, Zhang X, Liu M, Tian J, et al. Understanding LLMs: a comprehensive overview from training to inference. 2024. Preprint at arXiv: 2401.02038.

[34] Latif E, Zhou Y, Guo S, Gao Y, Shi L, Nayaaba M, et al. A systematic assessment of openAI o1-preview for higher order thinking in education. 2024. Preprint at arXiv: 2410.21287.

[35] Xiang L, Lin Z, Zhang L, Wu Z, Liu Z, Jiang H, et al. Artificial general intelligence for medical imaging analysis. IEEE Reviews in Biomedical Engineering. 2024.

[36] Wang P, Holmes J, Liu Z, Chen D, Liu T, Shen J, et al. A recent evaluation on the performance of LLMs on radiation oncology physics using questions of randomly shuffled options. 2024. Preprint at arXiv: 2412.10622.

[37] Ding Z, Liu Z, Jiang H, Gao Y, Zhai X, Liu T, et al. Foundation models for low-resource language education (vision paper). 2024. Preprint at arXiv: 2412.04774.

[38] Chen J, Peng S, Li Y, Zhao H, Jiang H, Pan Y, et al. Queen: a large language model for quechua-English translation. 2024. Preprint at arXiv: 2412.05184.

[39] Zhang Y, Pan Y, Zhong T, Dong P, Xie K, Liu Y, et al. Potential of multimodal large language models for data mining of medical images and free-text reports. Meta-Radiology. 2024;2(4):100103.

[40] Zhong T, Yang Z, Liu Z, Zhang R, Liu Y, Sun H, et al. Opportunities and challenges of large language models for low-resource languages in humanities research. 2024. Preprint at arXiv: 2412.04497.

[41] Jiang H, Pan Y, Chen J, Liu Z, Zhou Y, Peng S, et al. OracleSage: towards unified visual-linguistic understanding of oracle bone scripts through cross-modal knowledge fusion. 2024. Preprint at arXiv: 2411.17837.

[42] Liao W, Liu Z, Zhang Y, Huang X, Liu N, Liu T, et al. Zero-shot relation triplet extraction as next-sentence prediction. Knowl Base Syst. 2024;304:112507.

[43] Zhang L, Liu Z, Zhang L, Wu Z, Yu X, Holmes J, et al. Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology. Medical Physics. 2024;51(3):2187–99.

[44] Tan C, Cao Q, Li Y, Zhang J, Yang X, Zhao H, et al. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. 2023. Preprint at arXiv: 2312.17016.

[45] Shi Y, Peng S, Liu Z, Wu Z, Li Q, Xiang L. MGH radiology LLaMA: a LLaMA 3 70b model for radiology. 2024. Preprint at arXiv: 2408.11848.

[46] Peng S, Chen J, Liu Z, Wang H, Wu Z, Zhong T, et al. Transcending language boundaries: harnessing LLMs for low-resource language translation. 2024. Preprint at arXiv: 2411.11295.

[47] Wang J, Zhao H, Yang Z, Peng S, Chen J, Sun H, et al. Legal evalutions and challenges of large language models. 2024. Preprint at arXiv: 2411.10137.

[48] Holmes J, Zhang L, Ding Y, Feng H, Liu Z, Liu T, et al. Benchmarking a foundation large language model on its ability to relabel structure names in accordance with the American Association of Physicists in Medicine Task Group-263 report. Practical Radiation Oncology. 2024;14(6):e515–21.

[49] Lin Z, Wu Z, Dai H, Liu Z, Zhang T, Zhu D, et al. Embedding human brain function via transformer. In: International conference on medical image computing and computer-assisted intervention. Springer; 2022. p. 366–75.

[50] Lin Z, Wu Z, Dai H, Liu Z, Hu X, Zhang T, et al. A generic framework for embedding human brain function with temporally correlated autoencoder. Med Image Anal. 2023;89: 102892.

[51] Liu Z, Li Y, Zolotarevych O, Yang R, Liu T. LLM-POTUS score: a framework of analyzing presidential debates with large language models. 2024. Preprint at arXiv: 2409.08147.

[52] Yang Z, Lin X, He Q, Huang Z, Liu Z, Jiang H, et al. Examining the commitments and difficulties inherent in multimodal foundation models for street view imagery. 2024. Preprint at arXiv: 2408.12821.

[53] Gong X, Zhang J, Qi G, Teng Y, Hou J, Lyu Y, et al. Advancing microbial production through artificial intelligence-aided biology. Biotechnol Adv. 2024; 74:108399.

[54] Mukherjee S, Gamble P, Sanz Ausin M, Kant N, Aggarwal K, Manjunath N, et al. Polaris: a safety-focused LLM constellation architecture for healthcare. 2024. Preprint at arXiv: 2403.13313.

[55] Xu S, Wu Z, Zhao H, Shu P, Liu Z, Liao W, et al. Reasoning before comparison: LLM-enhanced semantic similarity metrics for domain specialized text analysis. 2024. Preprint at arXiv: 2402.11398.

[56] Latif E, Fang L, Ma P, Zhai X. Knowledge distillation of LLMs for automatic scoring of science assessments. In: International conference on artificial intelligence in education. Springer; 2024. p. 166–74.

[57] Liu Z, Holmes J, Liao W, Liu C, Zhang L, Feng H, et al. The radiation oncology NLP database. 2024. Preprint at arXiv: 2401.10995.

[58] Wei Y, Zhong T, Zhang S, Li X, Zhang T, Lin Z, et al. Chat2Brain: a method for mapping open-ended semantic queries to brain activation maps. In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2023. p. 1523–30.

[59] Liu Z, Jiang H, Zhong T, Wu Z, Ma C, Li Y, et al. Holistic evaluation of GPT-4V for biomedical imaging. 2023. Preprint at arXiv: 2312.05256.

[60] Liu Z, He M, Jiang Z, Wu Z, Dai H, Zhang L, et al. Survey on natural language processing in medical image analysis. Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences. 2022;47(8):981–93.

[61] Wu Z, Zhang L, Cao C, Yu X, Dai H, Ma C, et al. Exploring the trade-offs: unified large language models vs local fine-tuned models for highly-specific radiology NLI task. 2023. Preprint at arXiv: 2304.09138.

[62] Xiao Z, Chen Y, Zhang L, Yao J, Wu Z, Yu X, et al. Instruction-ViT: multi-modal prompts for instruction learning in ViT. 2023. Preprint at arXiv: 2305.00201.

[63] Cai H, Huang X, Liu Z, Liao W, Dai H, Wu Z, et al. Exploring multimodal approaches for Alzheimer's disease detection using patient speech transcript and audio data. 2023. Preprint at arXiv: 2307.02514.

[64] Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. 2023. Preprint at arXiv: 2304.01938.

[65] Liu Z, Wu Z, Hu M, Zhao B, Lin Z, Zhang T, et al. PharmacyGPT: the AI pharmacist. 2023. Preprint at arXiv: 2307.10432.

[66] Guan Z, Wu Z, Liu Z, Wu D, Ren H, Li Q, et al. CohortGPT: an enhanced GPT for participant recruitment in clinical study. 2023. Preprint at arXiv: 2307.11346.

[67] Liu Z, Zhong T, Li Y, Zhang Y, Pan Y, Zhao Z, et al. Evaluating large language models for radiology natural language processing. 2023. Preprint at arXiv: 2307.13693.

[68] Cai H, Huang X, Liu Z, Liao W, Dai H, Wu Z, et al. Multimodal approaches for Alzheimer's detection using patients' speech and transcript. In: International conference on brain informatics. Springer; 2023. p. 395–406.

[69] Shi Y, Xu S, Liu Z, Liu T, Xiang L, Liu N. MedEdit: model editing for medical question answering with external knowledge bases. 2023. Preprint at arXiv: 2309.16035.

[70] Tang C, Liu Z, Ma C, Wu Z, Li Y, Liu W, et al. PolicyGPT: automated analysis of privacy policies with large language models. 2023. Preprint at arXiv: 2309.10238.

[71] Liu Z, Li Y, Cao Q, Chen J, Yang T, Wu Z, et al. Transformation vs tradition: artificial general intelligence (AGI) for arts and humanities. 2023. Preprint at arXiv: 2310.19626.

[72] Zhong T, Zhao W, Zhang Y, Pan Y, Dong P, Jiang Z, et al. ChatRadio-valuer: a chat large language model for generalizable radiology report generation based on multi-institution and multi-system data. 2023. Preprint at arXiv: 2310.05242.

[73] Gong X, Holmes J, Li Y, Liu Z, Gan Q, Wu Z, et al. Evaluating the potential of leading large language models in reasoning biology questions. 2023. Preprint at arXiv: 2311.07582.

[74] Liao W, Liu Z, Zhang Y, Huang X, Qi F, Ding S, et al. Coarse-to-fine knowledge graph domain adaptation based on distantly-supervised iterative training. In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2023. p. 1294–9.

[75] Holmes J, Peng R, Li Y, Hu J, Liu Z, Wu Z, et al. Evaluating multiple large language models in pediatric ophthalmology. 2023. Preprint at arXiv: 2311.04368.

[76] Rezayi S, Liu Z, Wu Z, Chandra D, Ge B, Dai H, et al. Exploring new frontiers in agricultural NLP: investigating the potential of large language models for food applications. IEEE Transactions on Big Data. 2024;11(3):1235–46.

[77] Dou F, Ye J, Geng Y, Lu Q, Niu W, Sun H, et al. Towards artificial general intelligence (AGI) in the internet of things (IoT): opportunities and challenges. 2023. Preprint at arXiv: 2309.07438.

[78] Holmes J, Zhang L, Ding Y, Feng H, Liu Z, Liu T, et al. Benchmarking a foundation LLM on its ability to re-label structure names in accordance with the AAPM TG-263 report. 2023. Preprint at arXiv: 2310.03874.

[79] Sennrich R. Neural machine translation of rare words with subword units. 2015. Preprint at arXiv: 1508.07909.

[80] Schuster M, Nakajima K. Japanese and Korean voice search. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2012. p. 5149–52.

[81] Kudo T. Subword regularization: improving neural network translation models with multiple subword candidates. 2018. Preprint at arXiv: 1804.10959.

[82] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828.

[83] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.

[84] Lin Z, Feng M, dos Santos CN, Yu M, Xiang B, Zhou B, et al. A structured self-attentive sentence embedding. 2017. Preprint at arXiv: 1703.03130.

[85] Liu Z, Alavi A, Li M, Zhang X. Self-supervised contrastive learning for medical time series: a systematic review. Sensors. 2023;23(9):4221.

[86] Hastie T, Tibshirani R, Friedman J, Hastie T, Tibshirani R, Friedman J. Overview of supervised learning. In: The elements of statistical learning: data mining, inference, and prediction; 2009. p. 9–41.

[87] Geoffrey EH, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.

[88] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 1991;37(2):233–43.

[89] Vincent P, Larochelle H, Bengio Y. Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning; 2008. p. 1096–103.

[90] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol P-A, Bottou L. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res. 2010;11(12):3371–408.

[91] Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, et al. Language models as knowledge bases? 2019. Preprint at arXiv: 1909.01066.

[92] Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. Preprint at arXiv: 1801.06146.

[93] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.

[94] Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, et al. Pre-trained models: past, present and future. AI Open. 2021;2:225–50.

[95] Koumakis L. Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J. 2020;18:1466–73.

[96] Watson JD, Crick FH. On protein synthesis. In: The symposia of the society for experimental biology, 12; 1958. p. 138–63.

[97] Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, et al. From gene networks to gene function. Genome Res. 2003;13(12):2568–76.

[98] Kim CY, Baek S, Cha J, Yang S, Kim E, Marcotte EM, et al. HumanNet v3: an improved database of human gene networks for disease research. Nucleic Acids Res. 2022;50(D1):D632–9.

[99] Riad ABMKI, Abdul Barek M, Rahman MM, Akter MS, Islam T, Rahman MA, et al. Enhancing HIPAA compliance in AI-driven mHealth devices security and privacy. In: 2024 IEEE 48th annual computers, software, and applications conference (COMPSAC). IEEE; 2024. p. 2430–5.

[100] Bartels M. A balancing act: data protection compliance of artificial intelligence. GRUR Int. 2024;73(6):526–37.

[101] Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Comput Struct Biotechnol J. 2021;19:1750–8.

[102] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30.

[103] Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al. Simulating 500 million years of evolution with a language model. 2024. Preprint at bioRxiv: 2024.07.01.600583.

[104] Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43(11):1154–9.

[105] Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(Suppl l_1):D258–61.

[106] Ali A, Zhang J. Optimizing large language models: performance, efficiency, and scalability. East Eur J Multidiscip Res. 2024;3(2):13–9.

[107] Rostam ZRK, Szénási S, Kertész G. Achieving peak performance for large language models: a systematic review. IEEE Access. 2024;12:96017–50.

[108] Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artif Intell Rev. 2025;58(1):1–107.

[109] Naresh G, Thangavelu P. Integrating machine learning for health prediction and control in over-discharged li-nmc battery systems. Ionics. 2024;30(12):8015–32.

[110] Wu T, Luo L, Li Y-F, Pan S, Vu T-T, Haffari G. Continual learning for large language models: a survey. 2024. Preprint at arXiv: 2402.01364.

[111] Dohare S, Hernandez-Garcia JF, Lan Q, Rahman P, Mahmood AR, Sutton RS. Loss of plasticity in deep continual learning. Nature. 2024;632(8026):768–74.

[112] Bansal S, Sindhi V, Singla BS. Exploration of deep learning and transfer learning techniques in bioinformatics. In: Applying machine learning techniques to bioinformatics: few-shot and zero-shot methods. IGI Global; 2024. p. 238–57.

[113] Mishra SK, Singh A, Dubey KB, Kumar Paul P, Singh V. Role of bioinformatics in data mining and big data analysis. In: Advances in bioinformatics. Springer; 2024. p. 271–7.

[114] Yan B, Li K, Xu M, Dong Y, Zhang Y, Ren Z, et al. On protecting the data privacy of large language models (LLMs): a survey. 2024. Preprint at arXiv: 2403.05156.

[115] Zheng JY, Zhang HN, Wang LX, Qiu WJ, Zheng HW, Zheng ZM. Safely learning with private data: a federated learning framework for large language model. 2024. Preprint at arXiv: 2406.14898.

[116] Ye R, Wang W, Chai J, Li D, Li Z, Xu Y, et al. OpenFedLLM: training large language models on decentralized private data via federated learning. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining; 2024. p. 6137–47.

[117] Mao W, Hou C, Zhang T, Lin X, Tang K, Lv H. Parse trees guided LLM prompt compression. 2024. Preprint at arXiv: 2409.15395.

[118] Cai X, Wang C, Long Q, Zhou Y, Xiao M. Knowledge hierarchy guided biological-medical dataset distillation for domain LLM training. 2025. Preprint at arXiv: 2501.15108.

[119] Zhao H, Ma C, Xu FZ, Kong L, Deng Z-H. Biomaze: benchmarking and enhancing large language models for biological pathway reasoning. 2025. Preprint at arXiv: 2502.16660.

[120] Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller K-R. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE. 2021;109(3):247–78.

[121] Jia D, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. p. 248–55.

[122] Pourpanah F, Abdar M, Luo Y, Zhou X, Wang R, Lim CP, et al. A review of generalized zero-shot learning methods. IEEE Trans Pattern Anal Mach Intell. 2022;45(4):4051–70.

[123] Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv. 2020;53(3):1–34.

[124] Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. ACM J Data Inf Qual. 2023;15(2):1–21.

[125] Li H, Zhu C, Zhang Y, Sun Y, Shui Z, Kuang W, et al. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023. p. 7454–63.

[126] Zheng J, Hong H, Wang X, Su J, Liang Y, Wu S. Fine-tuning large language models for domain-specific machine translation. 2024. Preprint at arXiv: 2402.15061.

[127] Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. Adv Neural Inf Process Syst. 2024;36:34892–916.

[128] Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell. 2023;5(3):220–35.

[129] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. Lora: low-rank adaptation of large language models. 2021. Preprint at arXiv: 2106.09685.

[130] He R, Liu L, Ye H, Tan Q, Ding B, Cheng L, et al. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 2021. Preprint at arXiv: 2106.03164.

[131] Paul FC, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 2017;30.

[132] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. Preprint at arXiv: 1707.06347.

[133] Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. Adv Neural Inf Process Syst. 2024;36:53728–41.

[134] Amodei D, Olah C, Jacob S, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. 2016. Preprint at arXiv: 1606.06565.

[135] Pan A, Jones E, Jagadeesan M, Steinhardt J. Feedback loops with language models drive in-context reward hacking. 2024. Preprint at arXiv: 2402.06627.

[136] Hinton G. Distilling the knowledge in a neural network. 2015. Preprint at arXiv: 1503.02531.

[137] Xu X, Li M, Tao C, Shen T, Cheng R, Li J, et al. A survey on knowledge distillation of large language models. 2024. Preprint at arXiv: 2402.13116.

[138] Zhang H, Chen J, Jiang F, Yu F, Chen Z, Li J, et al. HuatuoGPT, towards taming language model to be a doctor. 2023. Preprint at arXiv: 2305.15075.

[139] Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Stanford Alpaca: an instruction-following LLaMA model. 2023.

[140] Hsieh C-Y, Li C-L, Yeh C-K, Nakhost H, Fujii Y, Ratner A, et al. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. 2023. Preprint at arXiv: 2305.02301.

[141] Liu A, Feng B, Xue B, Wang B, Wu B, Lu C, et al. DeepSeek-v3 technical report. 2024. Preprint at arXiv: 2412.19437.

[142] Firoozi R, Tucker J, Tian S, Majumdar A, Sun J, Liu W, et al. Foundation models in robotics: applications, challenges, and the future. Int J Robot Res. 2023;44(5):701–39.

[143] Liu J, Zhang C, Guo J, Zhang Y, Que H, Deng K, et al. DDK: distilling domain knowledge for efficient large language models. 2024. Preprint at arXiv: 2407.16154.

[144] Fang L, Chen Y, Zhong W, Ma P. Bayesian knowledge distillation: a Bayesian perspective of distillation with uncertainty quantification. In: Forty-first international conference on machine learning.

[145] Korattikara A, Rathod V, Murphy K, Welling M. Bayesian dark knowledge. 2015. Preprint at arXiv: 1506.04416.

[146] Touvron H, Lavril T, Izacard G, Martinet X, Jegou H, Grave E, et al. The LLaMA 3 herd of models. 2024. Preprint at arXiv: 2407.21783.

[147] Xu S, Zhou Y, Liu Z, Wu Z, Zhong T, Zhao H, et al. Towards next-generation medical agent: how o1 is reshaping decision-making in medical scenarios. 2024. Preprint at arXiv: 2411.14461.

[148] Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, et al. Opportunities and challenges for chatGPT and large language models in biomedicine and health. Briefings Bioinf. 2024;25(1):bbad493.

[149] Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015;31(3):405–12.

[150] Aly Abdelkader G, Kim J-D. Advances in protein-ligand binding affinity prediction via deep learning: a comprehensive study of datasets, data preprocessing techniques, and model architectures. Curr Drug Targets. 2024;25(15):1041–65.

[151] Kitts A, Sherry S. The single nucleotide polymorphism database (DBSNP) of nucleotide sequence variation. In: McEntyre J, Ostell JJ, editors. The NCBI handbook. Bethesda: US National Center for Biotechnology Information; 2002.

[152] Pal A, Kumar Umapathi L, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference, and learning. PMLR; 2022. p. 248–60.

[153] Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421.

[154] George T, Balikas G, Malakasiotis P, Partalas I, Zschunke M, Alvers MR, et al. An overview of the bioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinf. 2015;16:1–28.

[155] Nentidis A, Katsimpras G, Krithara A, Paliouras G. Overview of bioASQ tasks 12b and synergy12 in CLEF2024. In: Working notes of CLEF; 2024.

[156] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubmedQA: a dataset for biomedical research question answering. 2019. Preprint at arXiv: 1909.06146.

[157] Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng C-ying, Peng Y, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. 2019. Preprint at arXiv: 1901.07042.

[158] Wei C-H, Peng Y, Leaman R, Davis AP, Mattingly CJ, Jiao L, et al. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (CDR) task. Database. 2016;2016:baw032.

[159] Islamaj Doğan R, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inf. 2014;47:1–10.

[160] Taboureau O, Nielsen SK, Audouze K, Weinhold N, Edsgärd D, Roque FS, et al. ChemProt: a disease chemical biology database. Nucleic Acids Res. 2010;39(Suppl l_1): D367–72.

[161] Kim Kjærulff S, Wich L, Kringelum J, Jacobsen UP, Kouskoumvekaki I, Audouze K, et al. ChemProt-2.0: visual navigation in a disease chemical biology database. Nucleic Acids Res. 2012;41(D1):D464–9.

[162] Kringelum J, Kim Kjaerulff S, Brunak S, Lund O, Oprea TI, Taboureau O. ChemProt-3.0: a global chemical biology diseases mapping. Database. 2016;2016:bav123.

[163] Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inf. 2013;46(5):914–20.

[164] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinf. 2015;16:1–17.

[165] Smith L, Tanabe LK, Ando RJ, Kuo C-J, Chung I-F, Hsu C-N, et al. Overview of biocreative ii gene mention recognition. Genome Biol. 2008;9(S2):1–19.

[166] Collier N, Ohta T, Tsuruoka Y, Tateisi Y, Kim J-D. Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP); 2004. p. 73–8.

[167] Pustejovsky J, Castano J, Sauri R, Zhang J, Luo W. Medstract: creating large-scale information servers from biomedical texts. In: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain; 2002. p. 85–92.

[168] Gasperin C, Karamanis N, Seal R. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In: Proceedings of DAARC; 2007. Citeseer.

[169] Su J, Yang X, Hong H, Tateisi Y, Tsujii J. Coreference resolution in biomedical texts: a machine learning approach. Ont Text Min Life Sci. 2008;8.

[170] Segura-Bedmar I, Crespo M, de Pablo-Sánchez C, Martínez P. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. BMC Bioinf. 2010;11(S2):1–9.

[171] Nguyen N, Kim J-D, Tsujii J. Overview of bioNLP 2011 protein coreference shared task. In: Proceedings of BioNLP shared task 2011 workshop; 2011. p. 74–82.

[172] Theresa Batista-Navarro R, Ananiadou S. Building a coreference-annotated corpus from the domain of biochemistry. In: Proceedings of BioNLP 2011 workshop; 2011. p. 83–91.

[173] Bretonnel Cohen K, Lanfranchi A, Joo-young Choi M, Bada M, Baumgartner WA, Panteleyeva N, et al. Coreference annotation and resolution in the Colorado richly annotated full text (CRAFT) corpus of biomedical journal articles. BMC Bioinf. 2017;18:1–14.

[174] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4): 1234–40.

[175] Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist. 2020;8:64–77.

[176] Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics. 2016;32(3):432–40.

[177] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. 2018.

[178] Ji Y, Zhou Z, Liu H, Davuluri RV. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics. 2021; 37(15):2112–20.

[179] Sanabria M, Hirsch J, Poetsch AR. The human genome's vocabulary as proposed by the DNA language model GROVER. 2023. Preprint at bioRxiv: 2023.07.19.549677.

[180] Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y. Self-supervised learning on millions of pre-mRNA sequences improves sequence-based RNA splicing prediction. 2023. Preprint at bioRxiv: 2023.01.31.526427.

[181] Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. 2022. Preprint at arXiv: 2204.00300.

[182] Ahmed E, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. 2007. Preprint at arXiv: 2007.06225. arXiv 2020.

[183] Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun. 2022;13(1):4348.

[184] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 2022;13(1):1265.

[185] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):1–3.

[186] Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Henryk Grzywaczewski A, Oteri F, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nat Methods. 2024;22(2):1–11.

[187] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. DNABERT-2: efficient foundation model and benchmark for multi-species genome. 2023. Preprint at arXiv: 2306.15006.

[188] Nguyen E, Poli M, Faizi M, Thomas A, Wornow M, Birch-Sykes C, et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. Adv Neural Inf Process Syst. 2024;36.

[189] Poli M, Massaroli S, Nguyen E, Fu DY, Dao T, Baccus S, et al. Hyena hierarchy: towards larger convolutional language models. In: International conference on machine learning. PMLR; 2023. p. 28043–78.

[190] Zhang D, Zhang W, He B, Zhang J, Qin C, Yao J. DnaGPT: a generalized pretrained tool for multiple DNA sequence analysis tasks. 2023. Preprint at bioRxiv: 2023.07.11.548628.

[191] Zeng W, Gautam A, Huson DH. Mulan-methyl—multiple transformer-based language models for accurate DNA methylation prediction. GigaScience. 2023;12:giad054.

[192] Press O, Smith NA, Lewis M. Train short, test long: attention with linear biases enables input length extrapolation. 2021. Preprint at arXiv: 2108.12409.

[193] Dao T, Fu D, Ermon S, Rudra A, Flashattention CR. Fast and memory-efficient exact attention with IO-awareness. Adv Neural Inf Process Syst. 2022;35:16344–59.

[194] Zhou Z, Weimin W, Harrison H, Wang J, Lizhen S, Ramana VD, et al. DNABERT-S: pioneering species differentiation with species-aware DNA embeddings. 2024. Preprint at arXiv: 2402.08777.

[195] Dreos R, Ambrosini G, Cavin Périer R, Bucher P. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. Nucleic Acids Res. 2013;41(D1): D157–64.

[196] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414): 57–74.

[197] Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818): 699–710.

[198] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the encode project. Genome Res. 2012;22(9): 1760–74.

[199] Kalicki CH, Haritaoglu ED. RNABERT: RNA family classification and secondary structure prediction with BERT pretrained on RNA sequences.

[200] Chen K, Zhou Y, Ding M, Wang Y, Ren Z, Yang Y. Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. Briefings Bioinf. 2024;25(3):bbae163.

[201] Zhang Y, Lang M, Jiang J, Gao Z, Xu F, Litfin T, et al. Multiple sequence alignment-based RNA language model and its application to structural inference. Nucleic Acids Res. 2024;52(1):e3.

[202] Yamada K, Hamada M. Prediction of RNA–protein interactions using a nucleotide language model. Bioinform Adv. 2022;2(1):vbac023.

[203] Wright ES. RNAContest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency. RNA. 2020;26(5):531–40.

[204] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.

[205] Sweeney BA, Petrov AI, Ribas CE, Finn RD, Bateman A, Szymanski M, et al. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 2021;49(D1):D212–20.

[206] Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC genome browser database: 2019 update. Nucleic Acids Res. 2019;47(D1):D853–8.

[207] Pan X, Fang Y, Li X, Yang Y, Shen H-B. RBPsuite: RNA-protein binding sites prediction suite based on deep learning. BMC Genom. 2020;21:1–8.

[208] Zhang Q, Fan X, Wang Y, Sun M-an, Shao J, Guo D. BPP: a sequence-based algorithm for branch point prediction. Bioinformatics. 2017;33(20):3166–72.

[209] Scalzitti N, Kress A, Orhand R, Weber T, Moulinier L, Jeannin-Girardon A, et al. Spliceator: multi-species splice site prediction using convolutional neural networks. BMC Bioinf. 2021;22:1–26.

[210] Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10(1):5407.

[211] Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature. 2017;541(7637):331–8.

[212] Levine D, Asad Rizvi S, Lévy S, Pallikkavaliyaveetil N, Zhang D, Chen X, et al. Cell2sentence: teaching large language models the language of biology. 2023. Preprint at bioRxiv: 2023.09.11.557287.

[213] Shen H, Liu J, Hu J, Shen X, Zhang C, Wu D, et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. iScience. 2023;26(5):106536.

[214] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Sayed ZRA, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24.

[215] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1–11.

[216] Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell. 2022;4(10):852–66.

[217] Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. BMC Genom. 2019;20(S1):7–15.

[218] Xu J, Zhang A, Liu F, Chen L, Zhang X. CIForm as a transformer-based model for cell-type annotation of large-scale single-cell RNA-seq data. Briefings Bioinf. 2023;24(4): bbad195.

[219] Chen J, Xu H, Tao W, Chen Z, Zhao Y, Han J-DJ. Transformer for one stop interpretable cell type annotation. Nat Commun. 2023;14(1):223.

[220] Jiao L, Wang G, Dai H, Li X, Wang S, Song T. scTransSort: transformers for intelligent annotation of cell types by gene embeddings. Biomolecules. 2023;13(4):611.

[221] Song T, Dai H, Wang S, Wang G, Zhang X, Zhang Y, et al. Transcluster: a cell-type identification method for single-cell RNA-seq data using deep learning based on transformer. Front Genet. 2022;13:1038919.

[222] Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2023;24(1):21–43.

[223] Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196–203.

[224] Gao Z, Liu Q, Zeng W, Jiang R, Wong WH. EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics. Genome Biol. 2024;25(1): 1–30.

[225] Maxwell RM, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13(11):919–22.

[226] Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268–76.

[227] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.

[228] Steven RS. An introduction to mass spectrometry-based proteomics. J Proteome Res. 2023;22(7):2151–71.

[229] Wang F, Liu C, Li J, Yang F, Song J, Zang T, et al. SPDB: a comprehensive resource and knowledgebase for proteomic data at the single-cell resolution. Nucleic Acids Res. 2024;52(D1):D562–71.

[230] Ding N, Qu S, Xie L, Li Y, Liu Z, Zhang K, et al. Automating exploratory proteomics research via language models. 2024. Preprint at arXiv: 2411.03743.

[231] Zhang Q, Ding K, Lyv T, Wang X, Yin Q, Zhang Y, et al. Scientific large language models: a survey on biological & chemical domains. 2024. Preprint at arXiv: 2401.14656.

[232] Xiao H, Zhou F, Liu X, Liu T, Li Z, Liu X, et al. A comprehensive survey of large language models and multimodal large language models in medicine. 2024. Preprint at arXiv: 2405.08603.

[233] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci. 2021;118(15):e2016239118.

[234] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst. 2021;34:29287–303.

[235] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10.

[236] Ahmed E, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):7112–27.

[237] Ali M, McCann B, Naik N, Keskar NS, Anand N, Eguchi RR, et al. Progen: language modeling for protein generation. 2020. Preprint at arXiv: 2004.03497.

[238] Ferruz N, Schmidt S, Höcker B. A deep unsupervised language model for protein design. 2022. Preprint at bioRxiv: 2022.03.09.483666v1.

[239] Munsamy G, Lindner S, Lorenz P, Ferruz N. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes. In: NeurIPS machine learning in structural biology workshop; 2022.

[240] Hesslow D, Zanichelli N, Notin P, Poli I, Marks D. Rita: a study on scaling up generative protein sequence models. 2022. Preprint at arXiv: 2205.05789.

[241] Shuai RW, Ruffolo JA, Gray JJ. Generative language modeling for antibody design. 2021. Preprint at bioRxiv: 2021.12.13.472419v2.

[242] Sternke M, Karpiak J. ProteinRL: reinforcement learning with generative protein language models for property-directed sequence design. In: NeurIPS 2023 generative AI and biology (GenBio) workshop; 2023.

[243] Truong T, Jr., Bepler T. Poet: a generative model of protein families as sequences-of-sequences. Adv Neural Inf Process Syst. 2023;36:77379–415.

[244] Cao Y, Das P, Chenthamarakshan V, Chen P-Y, Melnyk I, Shen Y. Fold2Seq: a joint sequence (1D)-fold (3D) embedding-based generative model for protein design. In: International conference on machine learning. PMLR; 2021. p. 1261–71.

[245] Ram S, Bepler T. Few shot protein generation. 2022. Preprint at arXiv: 2204.01168.

[246] Sgarbossa D, Lupo U, Bitbol A-F. Generative power of a protein language model trained on multiple sequence alignments. eLife. 2023;12:e79854.

[247] Lee M, Felipe Vecchietti L, Jung H, Ro H, Cha M, Kim HM. Protein sequence design in a latent space via model-based reinforcement learning. 2023.

[248] Zheng Z, Deng Y, Xue D, Zhou Y, Ye F, Gu Q. Structure-informed language models are protein designers. In: International Conference on Machine Learning (ICML). ICML; 2023.

[249] Zhang L, Chen J, Shen T, Li Y, Sun S. Enhancing the protein tertiary structure prediction by multiple sequence alignment generation. 2023. Preprint at arXiv: 2306.01824.

[250] Heinzinger M, Weissenow K, Gomez Sanchez J, Henkel A, Steinegger M, Rost B. Bilingual language model for protein sequence and structure. NAR Genomics and Bioinformatics. 2024;6(4):lqae150.

[251] Chen B, Cheng X, Li P, Geng Y-ao, Gong J, Li S, et al. xTrimoPGLM: unified 100b-scale pre-trained transformer for deciphering the language of protein. 2024. Preprint at arXiv: 2401.06199.

[252] Serrano Y, Roda S, Guallar V, Molina A. Efficient and accurate sequence generation with small-scale protein language models. 2023. Preprint at bioRxiv: 2023.08.04.551626v1.

[253] Simon KSC, Wei KY. Generative antibody design for complementary chain pairing sequences through encoder-decoder language model. 2023. Preprint at arXiv: 2301.02748.

[254] Lee Y, Yu H, Lee J, Kim J. Pre-training sequence, structure, and surface features for comprehensive protein representation learning. In: The twelfth international conference on learning representations; 2023.

[255] Nguyen VTD, Son Hy T. Multimodal pretraining for unsupervised protein representation learning. Biol Methods Protoc. 2024;9(1):bpae043.

[256] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. 2022. Preprint at bioRxiv: 2022.07.20.500902v1.

[257] Durham J, Zhang J, Humphreys IR, Pei J, Cong Q. Recent advances in predicting and modeling protein–protein interactions. Trends Biochem Sci. 2023;48(6):527–38.

[258] AlQuraishi M. AlphaFold at CASP13. Bioinformatics. 2019;35(22):4862–5.

[259] Agarwal V, McShan AC. The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. Nat Chem Biol. 2024;20(8):950–9.

[260] Jha K, Karmakar S, Saha S. Graph-BERT and language model-based framework for protein–protein interaction identification. Sci Rep. 2023;13(1):5663.

[261] Li X, Han P, Chen W, Gao C, Wang S, Song T, et al. MARPPI: boosting prediction of protein–protein interactions with multi-scale architecture residual network. Briefings Bioinf. 2023; 24(1):bbac524.

[262] Lee JM, Hammarén HM, Savitski MM, Baek SH. Control of protein stability by post-translational modifications. Nat Commun. 2023;14(1):201.

[263] Shrestha P, Kandel J, Tayara H, Chong KT. Post-translational modification prediction via prompt-based fine-tuning of a GPT-2 model. Nat Commun. 2024;15(1):6699.

[264] Esmaili F, Pourmirzaei M, Ramazi S, Shojaeilangari S, Yavari E. A review of machine learning and algorithmic methods for protein phosphorylation site prediction. Genom Proteom Bioinform. 2023;21(6):1266–85.

[265] Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: an overview of protein structure prediction. Front Bioinform. 2023;3:1120370.

[266] Kim G, Lee S, Levy Karin E, Kim H, Moriwaki Y, Ovchinnikov S, et al. Easy and accurate protein structure prediction using colabFold. Nat Protoc. 2024;20(3):1–23.

[267] Jing B, Erives E, Pao-Huang P, Corso G, Berger B, Jaakkola T. EigenFold: generative protein structure prediction with diffusion models. 2023. Preprint at arXiv: 2304.02198.

[268] Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31.

[269] Nguyen T, Wriggers W, He J. A data set of paired structural segments between protein data bank and AlphaFold DB for medium-resolution cryo-em density maps: a gap in overall structural quality. In: International symposium on bioinformatics research and applications. Springer; 2024. p. 52–63.

[270] Lee GH, Min CW, Jang JW, Gupta R, Kim ST. Dataset on post-translational modifications proteome analysis of msp1-overexpressing rice leaf proteins. Data Brief. 2023;50:109573.

[271] Polina L, Weindl D. Dynamic models for metabolomics data integration. Curr Opin Syst Biol. 2021;28:100358.

[272] Tian L, Yu T. An integrated deep learning framework for the interpretation of untargeted metabolomics data. Briefings Bioinf. 2023;24(4):bbad244.

[273] Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. 2023. Preprint at arXiv: 2307.10169.

[274] Vu T, Siemek P, Bhinderwala F, Xu Y, Powers R. Evaluation of multivariate classification models for analyzing NMR metabolomics data. J Proteome Res. 2019;18(9):3282–94.

[275] Mao C, Xu J, Rasmussen L, Li Y, Adekkanattu P, Pacheco J, et al. AD-BERT: using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. J Biomed Inf. 2023;144:104442.

[276] Feng Y, Xu X, Zhuang Y, Zhang M. Large language models improve Alzheimer's disease diagnosis using multi-modality data. In: 2023 IEEE international conference on medical artificial intelligence (MedAI). IEEE; 2023. p. 61–6.

[277] Xie K, Gallagher RS, Conrad EC, Garrick CO, Baldassano SN, Bernabei JM, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. J Am Med Inf Assoc. 2022;29(5): 873–81.

[278] Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. 2024;34(3): e13207.

[279] Le Guellec B, Lefèvre A, Geay C, Shorten L, Bruge C, Hacein-Bey L, et al. Performance of an open-source large language model in extracting information from free-text radiology reports. Radiology: Artif Intell. 2024;6(4):e230364.

[280] Kanzawa J, Yasaka K, Fujita N, Fujiwara S, Abe O. Automated classification of brain MRI reports using fine-tuned large language models. Neuroradiology. 2024;66(12):1–7.

[281] Valsaraj A, Madala I, Garg N, Baths V. Alzheimer's dementia detection using acoustic & linguistic features and pre-trained BERT. In: 2021 8th international conference on soft computing & machine intelligence (ISCMI). IEEE; 2021. p. 171–5.

[282] Anand Vats N, Yadavalli A, Gurugubelli K, Kumar Vuppala A. Acoustic features, BERT model and their complementary nature for Alzheimer's dementia detection. In: Proceedings of the 2021 thirteenth international conference on contemporary computing; 2021. p. 267–72.

[283] Bang J-U, Han S-H, Kang B-O. Alzheimer's disease recognition from spontaneous speech using large language models. ETRI J. 2024;46(1):96–105.

[284] Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLoS Digit Health. 2022;1(12):e0000168.

[285] Cong Y, LaCroix AN, Lee J. Clinical efficacy of pre-trained large language models through the lens of aphasia. Sci Rep. 2024;14(1):15573.

[286] Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134–42.

[287] Lee J-H, Choi E, McDougal R, Lytton WW. GPT-4 performance for neurologic localization. Neurol Clin Pract. 2024;14(3): e200293.

[288] Kwon T, Tzu-iunn Ong K, Kang D, Moon S, Lee JR, Hwang D, et al. Large language models are clinical reasoners: reasoning-aware diagnosis framework with prompt-generated rationales. Proc AAAI Conf Artif Intell. 2024;38(16):18417–25.

[289] Akiyama M, Sakakibara Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. NAR Genom Bioinform. 2022;4(1):lqac012.

[290] Xu M, Yuan X, Miret S, Tang J. ProtST: multi-modality learning of protein sequences and biomedical texts. In: International conference on machine learning. PMLR; 2023. p. 38749–67.

[291] Liu Q, Zeng W, Zhu H, Li L, Wong WH, Alzheimer's Disease Neuroimaging Initiative. Leveraging genomic large language models to enhance causal genotype-brain-clinical pathways in Alzheimer's disease. 2024. Preprint at medRxiv 2024.10.03. 24314824.

[292] Frank M, Ni P, Jensen M, Gerstein MB. Leveraging a large language model to predict protein phase transition: a physical, multiscale, and interpretable approach. Proc Natl Acad Sci. 2024;121(33):e2320510121.

[293] Kim JW, Ahmed A, Bernardo D. EEG-GPT: exploring capabilities of large language models for EEG classification and interpretation. 2024. Preprint at arXiv: 2401.18006.

[294] Liu M, Song Z, Chen D, Wang X, Zhuang Z, Fei M, et al. Affinity learning based brain function representation for disease diagnosis. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024. p. 14–23.

[295] Ali B, Hashemi F. Brain-mamba: encoding brain activity via selective state space models. In: Conference on health, inference, and learning. PMLR; 2024. p. 233–50.

[296] Jan Holwerda T, Deeg DJH, Beekman ATF, Van Tilburg TG, Stek ML, Jonker C, et al. Feelings of loneliness, but not social isolation, predict dementia onset: results from the Amsterdam Study of the Elderly (AMSTEL). J Neurol Neurosurg Psychiatr. 2014;85(2):135–42.

[297] Xiang Q. ChatGPT: a promising tool to combat social isolation and loneliness in older adults with mild cognitive impairment. Neurol Live. 2023:NA.

[298] Raile P. The usefulness of chatGPT for psychotherapists and patients. Humanit Soc Sci Commun. 2024;11(1):1–8.

[299] Ali Mohammed I, Venkataraman S. An innovative study for the development of a wearable AI device to monitor Parkinson's disease using generative AI and LLM techniques. Int J Creat Res Thoughts. 2023:2320–882.

[300] Binta Manir S, Islam KMS, Madiraju P, Deshpande P. LLM-based text prediction and question answer models for aphasia speech. IEEE Access. 2024;12:114670–80.

[301] Liberati G, Rocha JLDD, Van der Heiden L, Raffone A, Birbaumer N, Belardinelli MO, et al. Toward a brain-computer interface for Alzheimer's disease patients by combining classical conditioning and brain state classification. J Alzheim Dis. 2012;31(s3):S211–20.

[302] Miladinović A, Ajčević M, Busan P, Jarmolowska J, Silveri G, Deodato M, et al. Evaluation of motor imagery-based BCI methods in neurorehabilitation of Parkinson's disease patients. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE; 2020. p. 3058–61.

[303] Li Z, Zhao S, Duan J, Su C-Y, Yang C, Zhao X. Human cooperative wheelchair with brain–machine interaction based on shared control strategy. IEEE ASME Trans Mechatron. 2016;22(1):185–95.

[304] Cao Z. A review of artificial intelligence for EEG-based brain-computer interfaces and applications. Brain Sci Adv. 2020;6(3):162–70.

[305] Sorino P, Biancofiore GM, Lofù D, Colafiglio T, Lombardi A, Narducci F, et al. ARIEL: brain-computer interfaces meet large language models for emotional support conversation. In: Adjunct proceedings of the 32nd ACM conference on user modeling, adaptation and personalization; 2024. p. 601–9.

[306] Jiménez Benetó DM. Arithmetic reasoning in large language models and a speech brain-computer interface. B.S. thesis. Universitat Politècnica de Catalunya; 2024.

[307] Reza Saeidnia H, Kozak M, Brady DL, Hassanzadeh M. Evaluation of chatGPT's responses to information needs and information seeking of dementia patients. Sci Rep. 2024;14(1):10273.

[308] Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res. 2023;25:e48966.

[309] Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inf Assoc. 2024;31(9):1833–43.

[310] Oh Y, Park S, Byun HK, Cho Y, Lee IJ, Kim JS, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun. 2024;15(1):9186.

[311] Oh Y, Park S, Xiang L, Yi W, Paly J, Efstathiou J, et al. Mixture of multicenter experts in multimodal generative AI for advanced radiotherapy target delineation. 2024. Preprint at arXiv: 2410.00046.

[312] Wang P, Liu Z, Li Y, Holmes J, Shu P, Zhang L, et al. Fine-tuning large language models for radiation oncology, a specialized health care domain. Int J Radiat Oncol Biol Phys. 2024;120(2):e664.

[313] Oh Y, Park S, Byun HK, Cho Y, Lee IJ, Kim JS, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun. 2024;15(1):9186.

[314] Dong Z, Chen Y, Gay H, Yao H, Hugo GD, Samson P, et al. Large-language-model empowered 3d dose prediction for intensity-modulated radiotherapy. Med Phys. 2024;52(1):619–32.

[315] Yuexing H, Holmes JM, Hobson J, Bennett A, Ebner DK, Routman DM, et al. Retrospective comparative analysis of prostate cancer in-basket messages: responses from closed-domain LLM vs. clinical teams. 2024. Preprint at arXiv: 2409.18290.

[316] Wang P, Holmes J, Liu Z, Chen D, Liu T, Shen J, et al. A recent evaluation on the performance of LLMs on radiation oncology physics using questions of randomly shuffled options. Front Oncol. 2025;15:1557064.

[317] Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. Prediction of influenza vaccination outcome by neural networks and logistic regression. J Biomed Inf. 2010;43(5):774–81.

[318] Alhasan K, Al-Tawfiq J, Aljamaan F, Jamal A, Al-Eyadhy A, Temsah M-H. Mitigating the burden of severe pediatric respiratory viruses in the post-Covid-19 era: ChatGPT insights and recommendations. Cureus. 2023;15(3):e36263.

[319] Hung S-K, Wu C-C, Singh A, Li J-H, Lee C, Chou EH, et al. Developing and validating clinical features-based machine learning algorithms to predict influenza infection in influenza-like illness patients. Biomed J. 2023;46(5):100561.

[320] Subba B, Toufiq M, Omi F, Yurieva M, Khan T, Rinchai D, et al. Large language model-driven selection of glutathione peroxidase 4 as a candidate blood transcriptional biomarker for circulating erythroid cells. 2024.

[321] Du J, Xiang Y, Sankaranarayanapillai M, Zhang M, Wang J, Si Y, et al. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. J Am Med Inf Assoc. 2021;28(7):1393–400.

[322] Shah SAW, Palomar DP, Barr I, Poon LLM, Ahmed AQ, McKay MR. Seasonal antigenic prediction of influenza A H3N2 using machine learning. Nat Commun. 2024;15(1):3833.

[323] Wu H, Li M, Zhang L. Comparing physician and large language model responses to influenza patient questions in the online health community. Int J Med Inf. 2025;197:105836.

[324] Huang X, Smith MC, Jamison AM, Broniatowski DA, Dredze M, Quinn SC, et al. Can online self-reports assist in real-time identification of influenza vaccination uptake? A cross-sectional study of influenza vaccine-related tweets in the USA, 2013–2017. BMJ Open. 2019;9(1):e024018.

[325] Li Y, Li J, He J, Tao C. AE-GPT: using large language models to extract adverse events from surveillance reports-a use case with influenza vaccine adverse events. PLoS One. 2024;19(3):e0300919.

[326] Mohamed Ghazy R, Elkhadry SW, Abdel-Rahman S, Taha SHN, Youssef N, Elshabrawy A, et al. External validation of the parental attitude about childhood vaccination scale. Front Public Health. 2023;11:1146792.

[327] Sammut F, Suda D, Caruana MA, Bogolyubova O. COVID-19 vaccination attitudes across the European continent. Heliyon. 2023;9(8):e18903.

[328] Skyles TJ, Stevens HP, Davis SC, Obray AM, Miner DS, East MJ, et al. Comparison of predictive factors of flu vaccine uptake pre- and post-COVID-19 using the NIS-teen survey. Vaccines. 2024;12(10):1164.

[329] Ahmad ST, Lu H, Liu S, Lau A, Amin B, Dras M, et al. VaxGuard: a multi-generator, multi-type, and multi-role dataset for detecting LLM-generated vaccine misinformation. 2025. Preprint at arXiv: 2503.09103.

[330] Sun VH, Heemelaar JC, Hadzic I, Raghu VK, Wu C-Y, Zubiri L, et al. Enhancing precision in detecting severe immune-related adverse events: comparative analysis of large language models and international classification of disease codes in patient records. J Clin Oncol. 2024;42(35):4134–44.

[331] Boubnovski Martell M, Märtens K, Phillips L, Keitley D, Dermit M, Fauqueur J. A scalable LLM framework for therapeutic biomarker discovery: grounding Q/A generation in knowledge graphs and literature. In: ICLR 2025 workshop on machine learning for genomics explorations.

[332] McIlwain DR, Chen H, Rahil Z, Bidoki NH, Jiang S, Bjornson Z, et al. Human influenza virus challenge identifies cellular correlates of protection for oral vaccination. Cell Host Microbe. 2021;29(12):1828–37.e5.

[333] Hayati M, Sobkowiak B, Stockdale JE, Colijn C. Phylo-genetic identification of influenza virus candidates for seasonal vaccines. Sci Adv. 2023;9(44):eabp9185.

[334] Gao C, Wen F, Guan M, Hatuwal B, Li L, Praena B, et al. MAIVeSS: streamlined selection of antigenically matched, high-yield viruses for seasonal influenza vaccine production. Nat Commun. 2024;15(1):1128.

[335] Montin D, Santilli V, Beni A, Costagliola G, Martire B, Mastrototaro MF, et al. Towards personalized vaccines. Front Immunol. 2024;15:1436108.

[336] Lee EK, Tian H, Nakaya HI. Antigenicity prediction and vaccine recommendation of human influenza virus A (H3N2) using convolutional neural networks. Hum Vaccines Immunother. 2020;16(11):2690–708.

[337] Meaney C, Escobar M, Stukel TA, Austin PC, Jaakkimainen L. Comparison of methods for estimating temporal topic models from primary care clinical text data: retrospective closed cohort study. JMIR Med Inform. 2022;10(12):e40102.

[338] Valerio V, Rampakakis E, Zanos TP, Levy TJ, Shen HC, McDonald EG, et al. High frequency of COVID-19 vaccine hesitancy among canadians immunized for influenza: a cross-sectional survey. Vaccines. 2022;10(9):1514.

[339] Ng QX, Ng CX, Ong C, Lee DYX, Liew TM. Examining public messaging on influenza vaccine over social media: unsupervised deep learning of 235,261 twitter posts from 2017 to 2023. Vaccines. 2023;11(10):1518.

[340] Ng QX, Lee DYX, Ng CX, Yau CE, Lim YL, Liew TM. Examining the negative sentiments related to influenza vaccination from 2017 to 2022: an unsupervised deep learning analysis of 261,613 twitter posts. Vaccines. 2023;11(6):1018.

[341] Levi Y, Brandeau ML, Shmueli E, Yamin D. Prediction and detection of side effects severity following COVID-19 and influenza vaccinations: utilizing smartwatches and smartphones. Sci Rep. 2024;14(1):6012.

[342] Deady M, Hussein E, Cook K, Billings D, Pizarro J, Plotogea AA, et al. The food and drug administration biologics

effectiveness and safety initiative facilitates detection of vaccine administrations from unstructured data in medical records through natural language processing. Front Digit Health. 2021;3:777905.

[343] Zimmermann MT, Kennedy RB, Grill DE, Oberg AL, Goergen KM, Ovsyannikova IG, et al. Integration of immune cell populations, mRNA-Seq, and CpG methylation to better predict humoral immunity to influenza vaccination: dependence of mRNA-Seq/CpG methylation on immune cell populations. Front Immunol. 2017;8:445.

[344] Galvan D, Effting L, Cremasco H, Adam Conte-Junior C. Can socioeconomic, health, and safety data explain the spread of Covid-19 outbreak on Brazilian Federative Units? Int J Environ Res Publ Health. 2020;17(23):8921.

[345] Wooden SL, Koff WC. The Human Vaccines Project: towards a comprehensive understanding of the human immune response to immunization. Hum Vaccines Immunother. 2018;14(9):2214–6.

[346] Hadid A, McDonald EG, Cheng MP, Papenburg J, Libman M, Dixon PC, et al. The WE SENSE study protocol: a controlled, longitudinal clinical trial on the use of wearable sensors for early detection and tracking of viral respiratory tract infections. Contemp Clin Trials. 2023;128:107103.

[347] Pletz MW, Vestergaard Jensen A, Bahrs C, Davenport C, Rupp J, Witzenrath M, et al. Unmet needs in pneumonia research: a comprehensive approach by the CAPNETZ study group. Respir Res. 2022;23(1):239.

[348] Liu J, Niu Q, Nagai-Tanima M, Aoyama T. Understanding human papillomavirus vaccination hesitancy in Japan using social media: content analysis. J Med Internet Res. 2025;27:e68881.

[349] Berdigaliyev N, Aljofan M. An overview of drug discovery and development. Future Med Chem. 2020;12(10):939–47.

[350] Cummings J, Zhou Y, Lee G, Zhong K, Fonseca J, Cheng F. Alzheimer's disease drug development pipeline: 2024. Alzheimer's Dement: Transl Res Clin Interv. 2024;10(2):e12465. eprint.

[351] Anastasiia V, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673–85.

[352] Wang J, Xiao Y, Shang X, Peng J. Predicting drug–target binding affinity with cross-scale graph contrastive learning. Briefings Bioinf. 2024;25(1):bbad516.

[353] Frey NC, Soklaski R, Axelrod S, Samsi S, Gómez-Bombarelli R, Coley CW, et al. Neural scaling of deep chemical models. Nat Mach Intell. 2023;5(11):1297–305.

[354] Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, Leskovec J, et al. A foundation model for clinician-centered drug repurposing. Nat Med. 2024;30(12):3601–13.

[355] Singhal K, Azizi S, Tu T, Sara Mahdavi S, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.

[356] Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D1180–92.

[357] Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45(D1):D985–94.

[358] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.

[359] Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. Trends Pharmacol Sci. 2023;44(9):561–72.

[360] Savage N. Drug discovery companies are customizing ChatGPT: here's how. Nat Biotechnol. 2023;41(5):585–6.

[361] Sheikholeslami M, Mazrouei N, Gheisari Y, Fasihi A, Irajpour M, Ali M. DrugGen: advancing drug discovery with large language models and reinforcement learning feedback. 2024. arXiv: 2411.14157 [q-bio].

[362] Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525–35.

[363] Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570–8.

[364] ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Briefings Bioinf. 2023;24(4):bbad226.

[365] Chaves JMZ, Wang E, Tu T, Dhaval Vaishnav E, Lee B, Sara Mahdavi S, et al. Tx-LLM: a large language model for therapeutics. 2024. arXiv: 2406.06316 [cs].

[366] Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci USA. 2023;120(24):e2220778120.

[367] Yang Z, Liu J, Yang F, Zhang X, Zhang Q, Zhu X, et al. Advancing Drug-Target Interaction prediction with BERT and subsequence embedding. Comput Biol Chem. 2024;110:108058.

[368] Kalakoti Y, Yadav S, Sundar D. TransDTI: transformer-Based Language Models for estimating DTIs and building a drug recommendation workflow. ACS Omega. 2022;7(3): 2706–17.

[369] Luo Z, Wu W, Sun Q, Wang J. Accurate and transferable drug–target interaction prediction with drugLAMP. Bioinformatics. 2024;40(12):btae693.

[370] Bal R, Xiao Y, Wang W. PGraphDTA: improving drug target interaction prediction using protein language models and contact maps. 2024. arXiv: 2310.04017 [cs].

[371] Fan Q, Liu Y, Zhang S, Ning X, Xu C, Han W, et al. CGPDTA: an explainable transfer learning-based predictor with molecule substructure graph for drug-target binding affinity. J Comput Chem. 2025;46(1):e27538.

[372] Liang Y, Zhang R, Li Y, Huo M, Ma Z, Singh D, et al. Multi-modal large language model enables all-purpose prediction of drug mechanisms and properties. 2024. 2024.09.29.615524. Section: New Results.

[373] Ma T, Lin X, Li T, Li C, Chen L, Zhou P, et al. Y-Mol: a multiscale biomedical knowledge-guided large language model for drug development. 2024. arXiv: 2410.11550 [cs].

[374] Inoue Y, Song T, Fu T. DrugAgent: explainable drug repurposing agent with large language model-based reasoning. 2024. arXiv: 2408.13378 [cs].

[375] Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative toxicogenomics database (CTD): update 2021. Nucleic Acids Res. 2021;49(D1):D1138–43.

[376] Chen L, Fan Z, Chang J, Yang R, Hou H, Guo H, et al. Sequence-based drug design as a concept in computational drug design. Nat Commun. 2023;14(1):4217.

[377] Zhang S, Xie L. Protein language model-powered 3d ligand binding site prediction from protein sequence. In: NeurIPS 2023 AI for science workshop; 2023.

[378] Fang X, Wang F, Liu L, He J, Lin D, Xiang Y, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. Nat Mach Intell. 2023;5(10):1087–96.

[379] Chakraborty C, Bhattacharya M, Lee S-S. Artificial intelligence enabled chatGPT and large language models in drug target discovery, drug discovery, and development. Mol Ther Nucleic Acids. 2023;33:866–8.

[380] Shaker B, Ahmad S, Lee J, Jung C, Na D. In silico methods and tools for drug discovery. Comput Biol Med. 2021;137:104851.

[381] Sharma G, Thakur A. ChatGPT in drug discovery. 2023.

[382] Morris GM, Huey R, Olson AJ. Using autodock for ligand-receptor docking. Curr Protoc Bioinform. 2008;24(1):8–14.

[383] Liang Y, Zhang R, Zhang L, Xie P. DrugChat: towards enabling chatGPT-like capabilities on drug molecule graphs. 2023. Preprint at arXiv: 2309.03907.

[384] Shen C, Zhang X, Deng Y, Gao J, Wang D, Xu L, et al. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. J Med Chem. 2022;65(15):10691–706.

[385] Trott O, Olson AJ. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010; 31(2):455–61.

[386] Bao J, He X, Zhang JZH. DeepBSP—a machine learning method for accurate prediction of protein–ligand docking structures. J Chem Inf Model. 2021;61(5):2231–40.

[387] Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK. A geometric deep learning approach to predict binding conformations of bioactive molecules. Nat Mach Intell. 2021; 3(12):1033–9.

[388] Niu Z, Xiao X, Wu W, Cai Q, Jiang Y, Jin W, et al. PharmaBench: enhancing admet benchmarks with large language models. Sci Data. 2024;11(1):985.

[389] Gawande MS, Zade N, Kumar P, Gundewar S, Nayodhara Weerarathna I, Verma P. The role of artificial intelligence in pandemic responses: from epidemiological modeling to vaccine development. Mol biomed. 2025;6(1):1.

[390] Anderson LN, Hoyt CT, Zucker JD, McNaughton AD, Teuton JR, Karis K, et al. Computational tools and data integration to accelerate vaccine development: challenges, opportunities, and future directions. Front Immunol. 2025;16:1502484.

[391] Tomic A, Tomic I, Rosenberg-Hasson Y, Dekker CL, Maecker HT, Davis MM. SIMON, an automated machine learning system, reveals immune signatures of influenza vaccine responses. J Immunol. 2019;203(3):749–59.

[392] Hayawi K, Shahriar S, Alashwal H, Serhani MA. Generative AI and large language models: a new frontier in reverse vaccinology. Inform Med Unlocked. 2024;48:101533.

[393] Luciani LL, Miller LM, Zhai B, Clarke K, Kramer KH, Schratz LJ, et al. Blood inflammatory biomarkers differentiate inpatient and outpatient coronavirus disease 2019 from influenza. Open Forum Infect Dis. 2023;10(3):ofad095.

[394] McGarvey PB, Suzek BE, Baraniuk JN, Rao S, Conkright B, Lababidi S, et al. In silico analysis of autoimmune diseases and genetic relationships to vaccination against infectious diseases. BMC Immunol. 2014;15(1):61.

[395] Kim M, Kim YJ, Park SJ, Kim KG, Oh PC, Kim YS, et al. Machine learning models to identify low adherence to influenza vaccination among Korean adults with cardiovascular disease. BMC Cardiovasc Disord. 2021;21(1):129.

[396] Cotugno N, Santilli V, Pascucci GR, Manno EC, De Armas L, Pallikkuth S, et al. Artificial intelligence applied to in vitro gene expression testing (IVIGET) to predict trivalent inactivated influenza vaccine immunogenicity in HIV infected children. Front Immunol. 2020;11:559590.

[397] Furman D, Jojic V, Kidd B, Shen-Orr S, Price J, Jarrell J, et al. Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. Mol Syst Biol. 2013;9(1): 659.

[398] Ruga T, Vocaturo E, Zumpano E. On the role of LLM to forecast the next pandemic. In: 2024 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE; 2024. p. 6567–73.

[399] Raj Saxena R. Examining reactions about Covid-19 vaccines: a systematic review of studies utilizing deep learning for sentiment analysis. 2024. Authorea Preprints.

[400] Hou AB, Du H, Wang Y, Zhang J, Wang Z, Liang PP, et al. Can a society of generative agents simulate human behavior and inform public health policy? A case study on vaccine hesitancy. 2025. Preprint at arXiv: 2503.09639.

[401] Sutskever I. Sequence to sequence learning with neural networks. 2014. Preprint at arXiv: 1409.3215.

How to cite this article: Ruan W, Lyu Y, Zhang J, Cai J, Shu P, Ge Y, et al. Large language models for bioinformatics. Quantitative Biology. 2026;e70014. https://doi.org/10.1002/qub2.70014