SCIENCE CHINA
Information Sciences

• REVIEW •

July 2025, Vol. 68, Iss. 7, 170101:1–170101:25
https://doi.org/10.1007/s11432-024-4466-3

Special Topic: AI for Biology

Large language models transform biological research: from architecture to utilization

Tao WANG[1,2*] & Zeyu LUO[3]

[1]*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[2]*Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an 710072, China*
[3]*College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China*

Received 8 January 2025/Revised 6 April 2025/Accepted 19 May 2025/Published online 19 June 2025

Abstract Recently, numerous large language models (LLMs) have emerged as foundational models, reshaping biological data modeling and achieving remarkable breakthroughs in both discriminative and generative tasks. The success of these models is largely attributed to the inherent similarities between natural language and biological data, such as DNA, RNA, and amino acid sequences. Through pre-training and fine-tuning phases, LLMs have demonstrated their ability to effectively model these biological datasets. Additionally, while protein structures and RNA-seq expression data are not inherently sequential, they can still be modeled and predicted effectively by LLMs based on the Transformer architecture. Previous research has predominantly focused on architectural innovations in LLMs and their applications to sequential data across various domains. However, there is a notable lack of systematic reviews addressing the reasons and methods behind LLM modifications for fitting biological omics data, particularly for non-sequential data types. Furthermore, comprehensive analyses of LLM applications in synthetic biology remain limited. We first systematically review representative LLMs in the biological domain. Next, we delve into their applications across the genome, transcriptome, and proteome fields, detailing the goals, processes, datasets, and methodologies involved. Finally, we discuss the challenges of applying LLMs to biological omics data and fundamental scientific research. In summary, we aim to provide a comprehensive overview of the technical and conceptual advances in this field, as well as an essential resource for researchers exploring the diverse applications of LLMs across various biological disciplines.

Keywords biological sequence and multimodal data, large language model, genome, transcriptome, proteome

Citation Wang T, Luo Z Y. Large language models transform biological research: from architecture to utilization. Sci China Inf Sci, 2025, 68(7): 170101, https://doi.org/10.1007/s11432-024-4466-3

1 Introduction

Large language models, a groundbreaking advancement in the field of natural language processing (NLP), have revolutionized global innovation with prominent examples such as ChatGPT [1,2] and Claude [3]. LLMs process sequential input data, such as sentences or paragraphs, and acquire their general-purpose, multitasking capabilities through a two-phase approach: pre-training and fine-tuning. At the heart of LLM architecture lies the Transformer framework [4], which forms the foundation of their design. Depending on the specific Transformer components utilized, LLMs are classified into three primary configurations: encoder-only, decoder-only, and encoder-decoder architectures. Representative models include BERT (bidirectional encoder representations from transformers) [5] for encoder-only designs, GPT (generative pre-trained transformer) for decoder-only models, and T5 (text-to-text transfer transformer) [6,7] for the encoder-decoder configuration. These sophisticated architectures enable LLMs to extract meaningful feature representations from raw, complex, and unstructured natural language sequences, effectively abstracting and decoding the underlying information embedded within text.

Just as natural language sequences encapsulate complex information, biological sequences and omics data store a wealth of information related to growth, development, genetics, and phenotypic regulation. These omics datasets, generated along the central dogma of molecular biology (DNA, RNA, protein), span genomics, transcriptomics, and proteomics, represent the most fundamental and information-rich layers of biological systems [8]. Developing models to construct high-quality representations and extract

* Corresponding author (email: twang@nwpu.edu.cn)

meaningful insights from biological omics data has remained a pivotal focus in advancing computational biology [9, 10], enabling a wide array of downstream applications.

The biological domain is increasingly embracing LLMs as cutting-edge tools for model-based analysis. Numerous domain-specific models have emerged, including Enformer [11], DNABERT [12], GROVER [13], the ESM family [14], and POET [15]. The success of these models stems from the inherent similarities between natural language and biological sequences [10, 12, 16], such as DNA, RNA, and amino acid sequences. This resemblance allows the development of biological LLMs using similar sequence modeling and pretraining strategies, including masked pretraining (BERT) [5, 17] and autoregressive pretraining (GPT) [1, 18]. Through fine-tuning, these LLMs can address fundamental biological challenges, such as predicting chromatin accessibility sites [19, 20], methylation modifications [21], protein subcellular localization [22, 23], protein-protein interaction sites [24], B-cell receptor specificity [25], protein 3D structure generation [17], and protein sequence-based remote homology searches [26].

While some types of biological omics data, such as structural data [27, 28] and RNA-seq expression data [18, 29, 30], are not inherently sequential, they can still be effectively modeled and predicted using foundational LLM paradigms built on the Transformer architecture. For example, scBERT [30] and scGPT [18] process RNA-seq expression matrices as input using specialized tokenization techniques and self-supervised pretraining to create general-purpose models for gene expression profiles at the transcriptomics level. Furthermore, to enhance the applicability of LLMs in bioinformatics, researchers are actively modifying LLM architectures to accommodate the redundancy and modality complexity inherent in biological omics data.

To provide cross-disciplinary researchers with a comprehensive understanding of LLM development and its applications in the biological domain, this study first systematically reviews the architecture and representative LLMs in the field of biology. We then detail the latest advancements in LLM research, including their goals, methodologies, architectures, datasets, and cutting-edge applications, such as those in synthetic biology, across the genome, transcriptome, and proteome landscapes. Additionally, we explore innovations in LLMs designed to address the redundancy and modality complexity of biological data. Finally, we discuss how transformative technologies and existing challenges may reshape LLM frameworks, highlighting future directions for their application in biological omics and fundamental scientific research.

Difference from existing surveys. Although previous reviews [31–34] have explored the applications of pretrained LLMs in the biological and biomedical fields, spanning genomics, transcriptomics, and proteomics, the rapid advancements in large models for biological data present challenges for a single review to capture the latest pretrained or fine-tuned models comprehensively. To bridge this gap, we have carefully curated the most representative models in this domain, focusing first on the architectures used to model biological data and their evolution. We then provide a detailed analysis of architectural modifications aimed at incorporating multimodal and evolutionary information into LLMs. The aim is to help researchers understand why such model designs are necessary, fostering deeper insight into the future development of LLM algorithms tailored for biological data. Furthermore, we adopt a narrative framework centered on the central dogma of molecular biology, offering a systematic and comprehensive examination of the diverse data types encompassed by the central dogma and their integration into LLM-based applications. We also highlight cutting-edge developments in adjacent fields, with a particular focus on the latest breakthroughs in synthetic biology.

Contribution. This survey provides a comprehensive overview of the technical foundations and conceptual advancements in the field, aiming to serve as a crucial resource for researchers investigating the diverse applications of large language models across biological omics.

2 Brief introduction to biological LLM architecture

2.1 Foundation architecture of LLMs

The Transformer is a deep learning architecture originally developed for natural language processing, which has since become the foundational model for a wide range of tasks, including those in genomics and transcriptomics. At a high level, the Transformer architecture excels at modeling long-range dependencies and complex patterns within sequential data by leveraging a self-attention mechanism. Unlike recurrent neural networks (RNNs), which process sequences step-by-step, the Transformer processes input sequences in parallel, allowing for faster training and better scalability. This parallel processing is enabled by self-

attention, which allows the model to weigh the importance of each token in a sequence relative to all others. As a result, the Transformer can capture contextual relationships more effectively, making it particularly well-suited for modeling the intricate patterns found in biological sequences such as DNA, RNA, and proteins.

2.1.1 *Attention & transformer*

The state-of-the-art (SOTA) performance achieved by LLMs, currently a major research focus, is predominantly rooted in the Transformer architecture. A Transformer comprises two core components: the encoder and decoder blocks. These blocks integrate advanced technologies such as attention mechanisms, layer normalization, and skip connections. At the heart of the Transformer lies the scaled dot-product attention mechanism (1), which models the relationships between tokens in a sequence. This mechanism represents each token as a weighted sum of all other tokens in the sequence. Tokens with lower relevance are assigned weights close to zero, while tokens with higher relevance receive greater weights. This process mimics human visual attention, focusing selectively on the most pertinent information, thus earning the term "attention mechanism".

The scaled dot-product attention employs three key weight matrices, Q (query), K (key), and V (value), to model relationships and generate attention-weighted representations, enabling the Transformer to effectively capture dependencies within the input data:

$$\text{Attention}\,(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V, \tag{1}$$

where $\sqrt{d_k}$ is the dimension of the keys. $\sqrt{d_k}$ is the normalization factor, which can ensure numerical stability.

In NLP, the Transformer architecture has enabled the development of models with hundreds of millions of parameters, including BERT, T5, and GPT. The BERT model leverages the Transformer's encoder architecture and employs bidirectional modeling to generate contextual representations of sequences. In contrast, the GPT model is based on the Transformer's decoder architecture and utilizes an autoregressive approach to model sequences unidirectionally, processing them from start to end. T5 incorporates both the encoder and decoder components of the Transformer and reframes all NLP tasks as text-to-text problems, unifying the approach to a wide range of applications. A brief illustration of these three models is shown in Figure 1.

2.1.2 *Key elements of LLM architecture*

Tokenization. The first step in constructing a Transformer model is tokenization, where raw data are converted into tokens. This step is crucial for transforming non-sequential data into sequential formats, as tokens can easily be arranged into a sequence.

Embedding. The embedding layer (including both positional and token embeddings) is fundamental to the Transformer architecture. It represents a critical evolution of word embeddings in both NLP and biological language models. These embeddings serve as the first step in mapping tokens into a vector space, capturing basic information from the tokens and guiding the model to fit both sequential and non-sequential data. In models like BERT and T5 (encoder), which employ random masking for pre-training, "mask" tokens are introduced to replace a preset proportion of tokens in the sequence. In contrast, the embedding layer in GPT and T5 (decoder) follows the original structure of the Transformer architecture, where token order plays a key role.

Pre-training. A key phase in training LLMs involves fitting the model to large amounts of unlabeled data to reconstruct or predict the original data. This self-supervised pre-training helps LLMs compress unstructured data, extract effective representations, and develop strong generalization abilities. During pre-training, BERT focuses on reconstructing masked tokens, while GPT predicts the next token based on the preceding tokens. T5, on the other hand, employs a more complex pre-training process, which includes reconstructing large portions of corrupted text. This paradigm allows BERT to retain bidirectional encoding capabilities, while GPT encodes only past context. Therefore, BERT may be more advantageous for modeling bidirectional sequences, such as genetic and amino acid sequences.

Fine-tuning. Fine-tuning LLMs for downstream tasks is a critical step in applying LLMs to biological research. This process adapts the general knowledge embedded in pre-trained models (manifested as

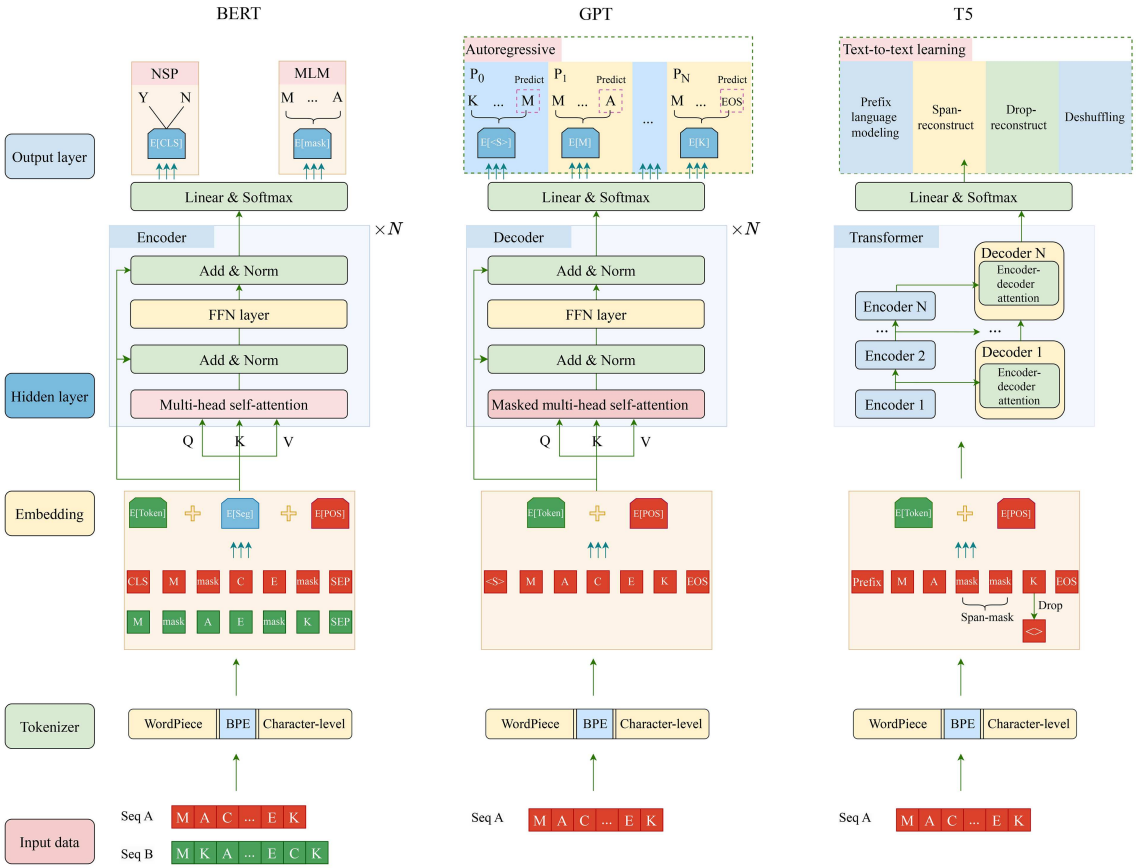

Figure 1 (Color online) Classical LLM architecture in the pre-training phase.

features and weights) to specific tasks, enhancing their suitability for targeted applications. Fine-tuning methods include feature extraction, full fine-tuning, parameter-efficient fine-tuning (PEFT) [35], instruction fine-tuning [36], and generating features from various modalities [24, 37, 38]. Feature extraction is widely used for its flexibility in building downstream models, while generative models for feature generation (e.g., structure-aware features) are becoming increasingly popular, particularly for fine-tuning models that incorporate 3D structural information. PEFT and instruction fine-tuning, which balance computational efficiency and performance, may become best practices for fine-tuning LLMs in the biological domain.

This section has covered the key elements of LLM architecture, including embedding, pre-training, and fine-tuning. Typically, LLMs and domain-specific LLMs are distinguished by their pre-training data: natural language for general LLMs and biological data for domain-specific models. In Subsection 2.2, we will focus on representative domain-specific LLMs, their frameworks, and training methods.

2.2 Representative models in the field of biology

Large language models are now being effectively applied in the biological and medical fields. To better adapt to domain-specific data, most biological domain LLMs are built upon the original LLM paradigm, with specialized modifications, as illustrated in Figure 2. In Table A1 in the supplementary file, we present the biological domain LLMs and downstream fine-tuning models discussed in this survey. Additionally, we provide a detailed comparison of several representative domain-specific LLMs, highlighting differences in their model architectures and pre-training approaches. To offer a comprehensive understanding of biological data, we also briefly present an illustration of the central dogma, encompassing the genome, transcriptome, and proteome.

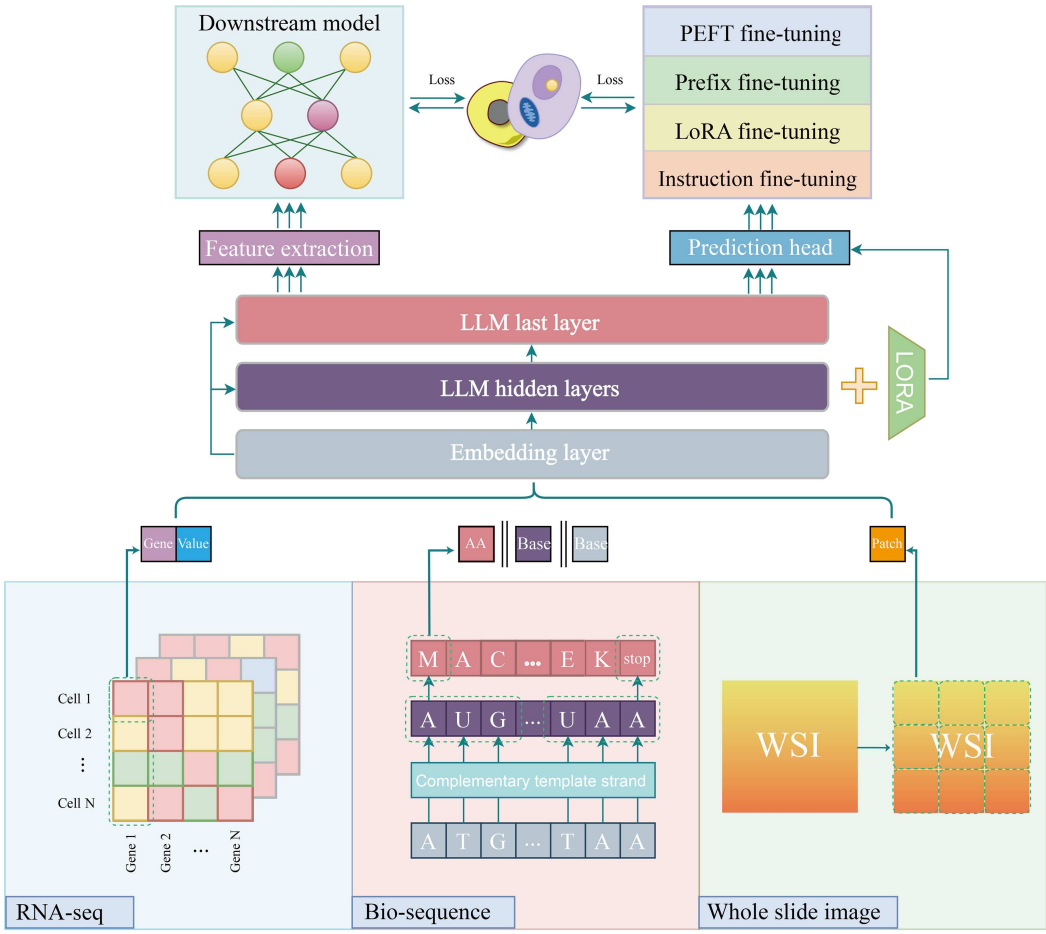

Figure 2 (Color online) Biological LLMs fine-tuning in downstream prediction. LLMs are demonstrated in fitting biological data, including RNA-seq, DNA, RNA, Amino acid sequence (protein primary structure), and biomedical image data (whole slide image, WSI). Each token corresponds to a gene base, an amino acid residue, and an image patch.

2.2.1 *Enformer versus DNABERT (genomics)*

Enformer [11] and DNABERT [12], both introduced in 2021, serve as foundational models for DNA sequence analysis. They excel at capturing long-range interactions within genes and have demonstrated strong modeling capabilities in non-coding regions. This is largely due to the attention mechanism, which enables the modeling of relationships across extended DNA sequences. The Enformer architecture combines a CNN block with Transformer encoders. The CNN block employs one-dimensional convolutions to process DNA sequences with fixed kernel lengths (window size), allowing for efficient feature extraction. In contrast, DNABERT follows the BERT paradigm for model construction, treating k-mers (k-th DNA base sequences) as tokens. These distinct approaches facilitate the downsampling of DNA sequences, making them more manageable for long-range modeling. Following these initial processes, the dimensional representation of DNA sequences can be described as follows:

$$N_{\text{Enformer}} = \frac{L}{S}, \tag{2}$$

where L is the original length of the DNA sequence, S is the stride of the convolution kernel, and N is the number of segments.

$$N_{\text{DNABERT}} = \left\lceil \frac{L}{K} \right\rceil, \tag{3}$$

where L is the original length of the DNA sequence, K is the size of the k-mer, and N is the length of all tokens in the sequence.

Additionally, DNABERT and its successor DNABERT2 follow the BERT-style masked pre-training approach, whereas Enformer employs a more traditional supervised learning method for pre-training.

Notably, DNABERT2 replaces k-mers with byte pair encoding (BPE) [39], which improves compression efficiency and reduces the length of DNA sequences. Gene sequences, in fact, are much longer than natural language sequences and more redundant than protein sequences. This highlights the importance of effective modeling techniques for genomic data, which could drive further research into downsampling methods for LLMs.

2.2.2 *scBERT versus scGPT (transcriptomics)*

RNA sequencing (RNA-seq) data are typically represented as a matrix $X \in \mathbb{R}^{C\times G}$, where C denotes the number of cells (in rows) and G denotes the number of genes (in columns). Each gene expression variable is non-negative, meaning $X \in R^{+}$. Although RNA-seq data are not inherently sequential, the complex interrelationships among genes within each cell are well-suited for modeling using the attention mechanisms in transformers. Therefore, to leverage the power of large language models (LLMs) for single-cell data, both scBERT and scGPT treat each gene as a token and consider the collection of all genes within a cell as a sequence. They replace traditional positional embeddings with expression embeddings, as the arrangement of genes in the feature columns has no inherent order.

Additionally, benefiting from the efficient language modeling paradigm, both scBERT and scGPT employ more extensive gene models and avoid aggressive filtering of high-variance genes (HVGs) during the pre-training phase to support a larger set of genes as tokens. For example, scBERT removes only cells with low counts of expressed genes without excessive gene filtering during data preprocessing, preserving over 15000 genes. scGPT extends this further by including the entire genome in the model. Moreover, since RNA-seq data are continuous values and often suffer from batch effects, both models utilize binning techniques in the expression embeddings to effectively transform row expression values into acceptable tokens. The key differences between scBERT and scGPT lie in their modeling architecture, particularly in the embedding layer, attention layer, and pre-training paradigms. In the embedding layer, scBERT includes only gene and expression tokens, whereas scGPT also incorporates condition tokens to represent gene attributes such as modality, batch, and perturbation conditions.

Additionally, scBERT introduces the concept of matrix factorization in the attention calculation, mapping the original Q and K matrices to Q' and K' through kernel transformation,

$$Q' = \phi\left(Q\right), K' = \phi\left(K\right), \tag{4}$$

where ϕ represents the kernel transformation function. This approach avoids the direct computation of QK^{T}, thereby alleviating the problem of excessively large QK^{T} matrices due to long sequence length. The modified module is known as the Performer [40].

Finally, scBERT follows BERT's masked language model (MLM) pre-training paradigm, whereas scGPT adopts a more complex approach. This complexity stems from the fact that autoregressive models are designed for sequences, but expression data do not inherently have a concept of sequence position, nor does it include a "next" gene in the traditional sense. To address this, scGPT innovatively organizes the autoregressive predictions based on attention scores. An interesting observation is that the designs of both models center around the attribution of expression data, emphasizing the importance of contextual relationships between genes within the single-cell environment.

2.2.3 *AlphaFold versus ESM (proteomics)*

In the field of proteomics, two prominent domain models have significantly advanced scientific understanding: the AlphaFold family introduced by DeepMind and the ESM family developed by Meta. These models represent two distinct approaches to constructing domain-specific models. Initially, AlphaFold2 [41] introduced the EvoFormer module, which includes the MSA (multiple sequence alignment) representation module to capture evolutionary information and the pair representation module to model spatial geometric data. This approach places a strong emphasis on incorporating domain knowledge to build predictive models.

In contrast, the ESM-1b [24] and its successor ESM2 [22] are based on a more general architecture rooted in the BERT model. ESM2 introduces rotary position embedding (RoPE) [42] in the embedding layer to handle longer protein sequences and updates the dynamic masking strategy, which is a key component of the RoBERTa model [43], during pre-training. As a result, ESM1b and ESM2 adhere more closely to the LLM paradigm. Interestingly, ESM2 has demonstrated superior accuracy in protein

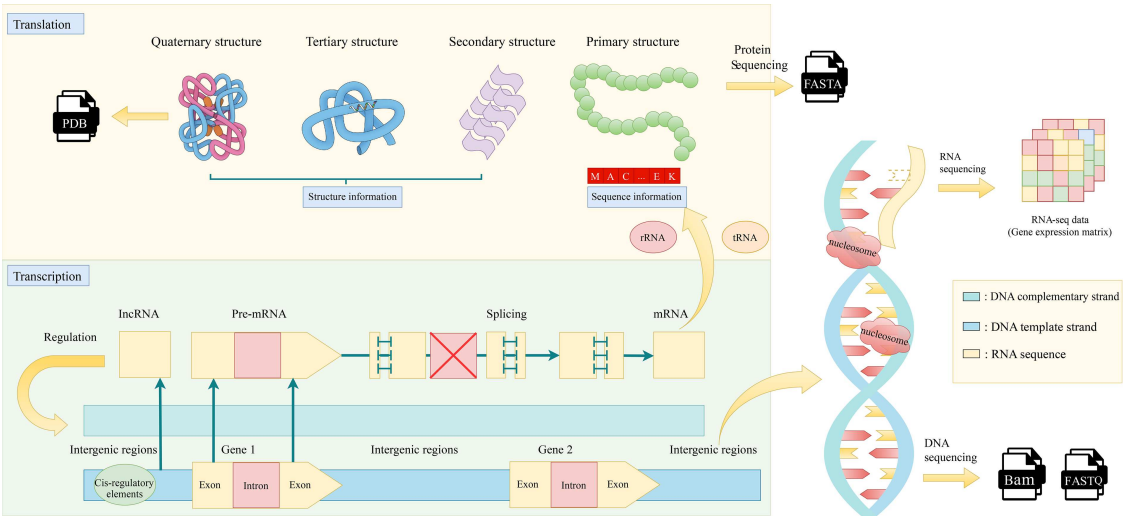

Figure 3 (Color online) Illustration of information flow in the central dogma of biology.

structure prediction compared to AlphaFold2, particularly in the absence of MSA information, which is one of AlphaFold2's major limitations.

For the latest versions, AlphaFold3 [44] and ESM3 [14], AlphaFold3 has simplified the EvoFormer module, reducing its reliance on the MSA module, and incorporates a diffusion architecture. This change leads to higher accuracy in generating protein-protein complex structures and enables the prediction of protein-molecular complex structures. ESM3 introduces structure and function tokens, capturing more biological prior knowledge, and employs geometric attention to better model the 3D structures of proteins. These innovations suggest that the two approaches are gradually converging. Notably, AlphaFold3 does not utilize an LLM architecture, whereas ESM3 remains firmly within the LLM framework. This raises an interesting question: Will AlphaFold4 adopt an LLM-based architecture, or will ESM4 continue to build on the LLM paradigm?

3 Applications of LLMs in genomics and transcriptomics field

3.1 Mission

In the central dogma of molecular biology (Figure 3), DNA encodes genetic information, which is then transcribed into RNA, reflecting gene expression at the transcriptional level. This genetic information is further translated into proteins, which manifest as biological functions at the translational level. With the advancement of sequencing technologies, vast amounts of data have emerged, shaping the fields of genomics, transcriptomics, and spatial omics [45–49]. A key distinction between genomic and transcriptomic data, as compared to proteomic data, is that the former consists of longer sequences with more redundancy, including non-coding regions and codon degeneracy, while also presenting more complex challenges related to sequencing quality [29, 45, 50, 51]. Like proteins, DNA and RNA also possess intricate spatial structures. Therefore, effectively embedding and representing genomic and transcriptomic data, encompassing DNA and RNA sequences, structural data, and expression profiles, remains a primary task for LLMs. This involves developing tokenizers and model architectures that are suited to the extended and diverse patterns inherent in genomic and transcriptomic data. Moreover, this field emphasizes research that spans from individual gene studies to broader genomic and transcriptomic levels.

The core focus areas in this field are summarized as follows.

(1) Developing optimized LLM models capable of fitting diverse genomic and transcriptomic data for downstream tasks.

(2) Extending research from gene function to broader genomic-level studies.

(3) Applying LLMs to the non-coding regions of sequences (e.g., intergenic regions in DNA, non-coding RNA).

(4) Advancing the use of LLMs in synthetic biology within the genomics and transcriptomics fields.

3.2 Algorithm and application

3.2.1 *Sequence modeling in DNA and RNA*

DNA sequences are inherently long and redundant, posing challenges for effective embedding in computational models. The primary objective is to develop methods for transforming raw sequence data into a tokenized format that can be effectively modeled by LLMs. This process begins with tokenization and embedding, where raw data are converted into tokens with vector embeddings. Some models achieve token embedding through specialized architectural designs. For example, the Enformer model applies one-dimensional convolutional kernels to embed DNA sequences, followed by processing in transformer modules. Although this method effectively models DNA sequences, it is limited by its reliance on a preset window size for the convolutional kernels, which constrains its ability to model DNA sequences of varying lengths—sequences longer than the fixed window size are truncated, affecting fine-tuning accuracy.

To address this, many LLMs adopt a more flexible Tokenizer paradigm, combining multiple bases into a single token to better accommodate DNA sequences of varying lengths. A common approach is k-mer encoding, where tokens are created by grouping bases (e.g., three bases at a time, or k-mers). This method is employed in models such as DNABERT [12] and DNAGPT [52]. Alternatively, subword tokenization techniques like WordPiece [53] and BytePairEncoding (BPE) [54] use statistical methods to iteratively generate tokens from different base combinations. Models such as DNABERT2 [55] and GROVER [13] adopt BPE-encoding for more flexible token generation. Subword tokenization offers an advantage because it generates variable-length base combinations, allowing LLMs to adapt better to the diversity of DNA sequences. However, the k-mer approach raises concerns about potential information leakage or sequence sensitivity, particularly when adjacent k-mer tokens share base overlaps. This issue is similar to codon degeneracy in nature, where a non-strict one-to-one mapping exists between three bases and an amino acid [45, 56]. Mapping bases directly to fixed tokens may overlook this flexibility. Just as nature employs a genetic "codebook" for translating genetic information, LLMs also need their own vocabulary rules for tokenization. Some models, like ENBED, opt for a single-base tokenization strategy, treating each DNA base as an individual token. This approach accounts for mutations such as single nucleotide polymorphisms (SNPs) [45], which can significantly alter model predictions. However, this method results in longer sequences, increasing the complexity of model training and inference.

RNA sequences, while similar to DNA in primary structure, differ by replacing thymine (T) with uracil (U). Many RNA-based models, such as UNI-RNA, convert all "U" bases in RNA sequences to "T" for consistency during training. Unlike DNA sequence tokenization, most RNA models (e.g., UNI-RNA, RNA-FM, RNA-MSM, ERNIE-RNA) use single-base tokenization. This is possible because RNA sequences are typically shorter than DNA sequences, representing individual units like mRNA or other small RNA molecules (e.g., tRNA). This makes single-base tokenization more suitable for RNA data. Similarly, protein sequences, which are shorter and less redundant than DNA, often use single amino acid residue tokenization in LLMs.

DNA and RNA sequences present challenges in terms of length, leading to improvements in LLM architectures. In particular, the scaled dot-product attention mechanism in transformers has a quadratic complexity in relation to sequence length ($\mathrm{O}(L^2)$, where L is the total number of tokens) [57]. To reduce this complexity, some models, like DNABERT2 and UNI-RNA, integrate the flash attention module, which improves the input/output (I/O) process during attention computations, thereby reducing complexity. Most classical transformer models rely on absolute positional encoding in the embedding layer, but this approach becomes problematic when fine-tuning sequences of varying lengths, as it cannot effectively handle longer sequences not encountered during pre-training. To address this, models like DNABERT2 replace traditional positional encodings with linear biases, which allow for more flexible handling of sequence length. UNI-RNA uses RoPE, a method also employed by the large protein language model ESM2, which facilitates the extension of sequence length beyond the limits imposed by traditional positional encoding.

3.2.2 *RNA-seq expression matrix in transcriptomics*

In transcriptomics, the expression matrix is essential for understanding gene expression and transcriptional activity, forming the foundation for various downstream analyses. The primary method for generating such data is RNA sequencing (RNA-seq), which includes both bulk RNA-seq at the tissue level and single-cell RNA sequencing (scRNA-seq) at the cellular level. Specifically, scRNA-seq provides high-

resolution data by sequencing individual cells, enabling more detailed insights into gene expression at the single-cell level. RNA-seq data, after undergoing upstream processing, is structured as an expression profile matrix. This structured format is essential for diverse analyses, including gene regulation studies [58], drug sensitivity prediction [59, 60], and disease prognosis [61]. Statistical models, including machine learning and deep learning approaches, have been widely applied to RNA-seq data before the integration of LLMs for modeling biological sequences. The matrix format of RNA-seq data makes it particularly suitable for traditional statistical models designed for tabular data, but the emergence of LLMs has opened new possibilities.

While traditional sequence-based LLMs are not inherently suited for modeling RNA-seq data due to the lack of sequence positional information, the attention mechanism in transformers can effectively capture complex relationships between genes. This ability to model intricate relationships, combined with the large-scale data integration potential of pretrained models, positions LLMs as valuable tools for RNA-seq data modeling. Notable early studies in this area include scBERT [30] and scGPT [18], which adapt the BERT and GPT architectures for modeling scRNA-seq data. These models include specific modifications to handle the nature of RNA-seq data during their pre-training phases. For example, scBERT uses the BERT architecture, while scGPT leverages GPT to adapt to the unique challenges of single-cell RNA-seq data. Another significant contribution is scFoundation [29], which employs an asymmetric encoder-decoder transformer architecture called xTrimoGene. This model uses a masked language model (MLM) training objective, where masked genes are reconstructed based on unmasked genes and two special overall gene expression vectors (T&S). This structure enables effective modeling for RNA-seq data with varying sequencing depths. Additionally, scFoundation incorporates the Performer architecture, replacing traditional attention mechanisms to reduce computational complexity, making it suitable for large-scale RNA-seq datasets.

During the fine-tuning phase, cell-level and gene-level tasks can be addressed by treating cells as sequences and individual genes as tokens. The fine-tuning tasks can be categorized as follows.

(1) Cell-level tasks, such as cell type annotation, cell trajectory analysis, drug response prediction, and single-cell perturbation prediction.

(2) Gene-level tasks, including drug sensitivity gene identification and key gene screening.

One advantage of models like scGPT is the inclusion of a CLS token as a special starting token, allowing the model to rely solely on features extracted from this token for cell-level tasks. This approach simplifies the fine-tuning process. In contrast, scBERT and xTrimoGene require the aggregation of features from all genes, typically through pooling or other methods, to form a cell representation for fine-tuning. This distinction highlights the different approaches to handling cell-level tasks and offers insight into the flexibility of LLMs for RNA-seq data analysis.

3.2.3 *Application in non-coding sequences*

Genomic data contains vast amounts of information, but when research focuses on the segments of DNA that encode proteins, specifically the exonic regions of genes, most of the genome, including introns and intergenic regions (non-coding areas), can be considered redundant. This is particularly true for the human genome and most eukaryotic genomes. However, the complex regulatory processes that govern the flow of genetic information from DNA to proteins cannot function without the involvement of these non-coding regions. While these areas are less well understood, contain more redundancy, and have fewer high-quality experimental labels (as they do not code for proteins and are often referred to as "junk DNA"), they play critical roles in gene regulation.

The scarcity of high-quality labels in non-coding regions makes traditional supervised learning approaches more challenging. However, self-supervised learning models, like LLMs, do not require extensive labeled data during pre-training and need fewer labels for fine-tuning. This characteristic makes them especially well-suited for exploring and modeling these non-coding regions. LLMs have proven effective in modeling and predicting the function of transcription factors, cis-regulatory elements, and promoters, all of which are located in non-coding regions of the genome. For example, models such as DNABERT-Prom-300 [12], DNABERT-Prom-scan [12], and miProBERT [62], which are fine-tuned versions of DNABERT [12], can accurately identify TATA and non-TATA promoters. Similarly, the DNABERT-TF [12] model, also fine-tuned from DNABERT, excels at predicting transcription factor binding sites (TFBS). These models perform exceptionally well on these tasks, especially when fine-tuned for specific datasets.

Furthermore, the fine-tuned DNABERT2 model achieves high performance in promoter detection (PD) and core promoter detection (CPD). However, its performance in human data may not surpass that of DNABERT due to potential limitations of byte-pair encoding (BPE) when encoding short DNA sequences. Models like GROVER and DNAGPT have also been successfully fine-tuned for predicting cis-regulatory regions, expanding the scope of DNA sequence modeling. In a noteworthy application, Chang et al. [63] demonstrated that a BERT-based multilingual model, originally pre-trained on cross-linguistic corpora, could be adapted to predict DNA sequences. After fine-tuning on DNA data, this model was used to predict whether a given DNA sequence belonged to a promoter region. This finding highlights the versatility of pre-trained language models, which can be fine-tuned for DNA sequence prediction tasks even when originally trained on natural language corpora. Models like PromoGen [64], based on the GPT-2 architecture, are pretrained on DNA promoter sequences, focusing on specific regions rather than the entire genome. This pretraining approach eliminates the need to consider sequence length constraints, as the model is designed to handle promoter sequences specifically. PromoGen is also fine-tuned using species-specific datasets to generate and design promoter sequences tailored to individual species.

3.2.4 *Application in epigenetics*

Epigenetics has become a central focus in life sciences research, examining how gene activity regulation changes without altering the underlying DNA sequence. This field includes processes such as DNA methylation, histone modification, and chromatin accessibility, all of which influence gene expression. As the field progresses, LLMs have increasingly been applied to support epigenetic research and modeling.

One notable example is the BERT6mA model [65], which is based on the BERT architecture and undergoes cross-species pre-training followed by fine-tuning for the target species to predict 6mA methylation sites. iDNA-ABF [21], fine-tuned from DNABERT, is capable of predicting various types of methylation sites, improving its performance by incorporating histone modification coverage information for more accurate detection of 5mC methylation sites. Another model, iDNA-ABT [66], also based on the BERT framework, shows high precision in detecting multiple methylation sites by integrating the CLS token in the tokenizer and treating each nucleotide as a distinct token. MuLan-Methyl [67] explores various BERT-related pre-training frameworks and processes, including masked language model pre-training (with the exception of ELECTRA [68], which uses a generator-discriminator approach to identify token replacements). MuLan-Methyl uses a WordPiece tokenizer and incorporates the CLS token to fine-tune and detect multiple methylation sites across different species.

Another promising model is EPiGePT, which is based on the Transformer encoder and combines MLM pre-training with multitask learning. This model can handle a range of epigenetic tasks, such as transcription factor binding, histone modification, and chromatin accessibility, without the need for additional fine-tuning after pre-training. EPiGePT's tokenizer approach for DNA sequences is similar to Enformer, involving a convolutional embedding process applied after one-hot encoding of sequences with a fixed length of 128 bases. Additionally, EPiGePT integrates transcription factor expression data (RNA-seq), which enhances the model's predictive power by incorporating multi-modal data, positioning it as a potential foundational model in epigenomics.

3.2.5 *Application at the genome level*

With advancements in high-scalability token embedding techniques in LLMs and their enhanced ability to capture long-range sequence relationships, there has been a significant expansion in the capacity to model and predict genomic features on a broad scale. This development parallels the shift from analyzing individual sentences or paragraphs to studying entire books, thus providing a comprehensive perspective on genomics that enables the study of organismal characteristics and functions at the genome level.

Notable research in this domain includes the following.

• Genome-wide mutation (or variant effect) prediction. LLMs have been utilized for predicting the impact of genetic mutations across the genome, which is crucial for understanding disease susceptibility and treatment responses.

• Genome-wide modification prediction. Predicting modifications across the entire genome, such as epigenetic changes or structural variations, helps in understanding genome stability and regulatory mechanisms.

• Identification of functional traits at the species genome level. LLMs have been applied to identify key functional traits, such as resistance mechanisms, in the genomes of different species, which can aid

in agriculture, medicine, and evolutionary studies.

These applications highlight the growing ability of LLMs to extend their predictive power beyond individual genes, encompassing the entire genome to provide insights into complex biological functions and characteristics.

3.2.6 *Application in synthetic biology*

Recent groundbreaking advancements in synthetic biology have focused on the de novo design and synthesis of functional genomes that can support cellular metabolism and self-replication [69–71]. Achieving these results requires a deep understanding of genome function, enhancing the efficiency of genome editing tools (such as the CRISPR-Cas9 system), and a comprehensive grasp of overall cellular functions. LLMs in genomics and transcriptomics have already facilitated several key applications in synthetic biology, including the following.

(1) Functional genome annotations at the genome scale [72, 73]. LLMs can provide insights into the functional elements of genomes, enabling the design of synthetic genomes with well-defined functions.

(2) RNA structure prediction and design [69, 74–76]. LLMs have proven effective in predicting and designing RNA structures, which is critical for synthetic biology, especially in the context of CRISPR-related applications.

(3) Gene editing system design [77]. LLMs are being used to optimize CRISPR-based gene editing systems, making them more precise and efficient.

(4) Single-cell functional annotations based on scRNA-seq data [18, 29, 30]. LLMs also support functional analysis at the single-cell level, which is crucial for creating personalized and highly specific biological systems in synthetic biology.

These advancements demonstrate the growing potential of LLMs to bridge genomics, transcriptomics, and synthetic biology, providing powerful tools for the design and optimization of synthetic biological systems.

3.2.7 *Integration of multimodal information*

In addition to primary sequence information, genomic and transcriptomic data often encompass other valuable modalities, including secondary structure, functional annotations, and evolutionary conservation [75]. Several recent models, such as DNAGPT, ERNIE-RNA, and RNA-MSM, have begun to incorporate these complementary data sources to enhance predictive performance. For example, DNAGPT integrates numerical features alongside DNA sequences to capture functional signals, while ERNIE-RNA introduces pairwise positional biases in its attention mechanism to infer RNA secondary structure from sequence data. RNA-MSM incorporates evolutionary information to enrich its representation learning and improve downstream predictions.

Although most current LLMs in genomics and transcriptomics focus primarily on sequence data, the growing availability of high-quality experimental and LLM-generated annotations paves the way for broader multimodal integration. This direction holds great promise for enhancing both predictive power and biological interpretability. By jointly learning from sequence data, expression profiles, structural features, and functional annotations, future models can develop more comprehensive representations that reflect not only the primary nucleotide sequence but also its regulatory, structural, and evolutionary context. This enables more accurate modeling of complex biological processes such as gene regulation, alternative splicing, and cellular state transitions across diverse conditions and species.

Moreover, multimodal integration can improve model robustness and generalizability, particularly for applications involving novel tasks or datasets with limited labeled data. Ultimately, such integration supports a more holistic understanding of biological function, aligning model predictions more closely with real-world biological systems and enhancing their utility in both basic research and translational contexts.

Despite these advantages, integrating multimodal information into LLMs presents notable challenges. The convergence of diverse data types, ranging from nucleotide sequences and expression profiles to protein structures, functional annotations, and evolutionary conservation, offers rich biological context but also introduces complexity. Aligning and encoding heterogeneous modalities into a unified model framework is nontrivial, especially given their varying resolutions, formats, and noise levels. Scalability is another concern, as multimodal LLMs require significant computational resources and sophisticated

architectures capable of processing modality-specific features while preserving interpretability. Additionally, the limited availability of fully aligned multimodal datasets and the need to effectively handle missing or partial modalities pose further barriers.

Addressing these challenges will be critical for realizing the full potential of multimodal LLMs in genomics and transcriptomics. Success in this area could yield generalizable, biologically grounded models that drive forward both fundamental discovery and clinical innovation.

3.3 Database

Most genomic datasets predominantly store DNA and RNA sequences, with the Human Genome Project Database being one of the most prominent genomic databases. In contrast, gene expression data are preserved in RNA sequencing databases, which are divided into bulk and single-cell RNA-seq data. Additionally, advanced research-related databases involved RNA's spatial structure, DNA methylation sequencing data (epigenetics related), and spatial transcriptomics data (RNA-seq related).

For pre-training purposes, large datasets are in the nature fit for this task due to their large scale. In comparison, smaller datasets need to be amalgamated with other datasets for pre-training. Theoretically, any dataset that provides ample data to train a large model could be suitable for pretraining, contingent primarily upon quality control measures and conventional selection practices. For fine-tuning, the scale of data is not a constraint, but comprehensive task-related annotations are essential. Table 1 [28, 33, 70, 73, 78–94] delineates commonly employed datasets for pretraining or fine-tuning LLMs in the domains of genomics and transcriptomics, highlighting the datasets' attributes and application scenarios.

3.4 Summary

This section delves into the application of LLMs in genomics and transcriptomics, emphasizing their purpose, the diverse model frameworks suited for various data types, and a comparison of tokenization methods for embedding long and complex sequences. Moreover, cutting-edge applications of LLMs across different domains of genomics and transcriptomics are discussed, alongside a summary of relevant datasets. With advancements in high-throughput sequencing technologies, the scope and format of sequencing data have significantly evolved, especially in areas such as non-coding regions, methylation sequencing [95], and spatial transcriptomics [96, 97]. As LLMs excel in handling large-scale and multimodal data, they hold immense potential for continued and widespread application in genomics and transcriptomics.

4 Application of LLMs in the proteomics field

4.1 Mission

Proteins, often regarded as the "final output" in the Central Dogma (see Figure 3), play a pivotal role in facilitating essential life functions. The functional basis of proteins stems from the folding of their tertiary spatial structure, with the instructions for this folding process encoded in their primary structure (amino acid sequences) [98]. Understanding protein functionality, encompassing aspects such as specific expression, subcellular localization [22, 23], and interactions with other molecules (e.g., protein-protein interactions [44], protein-ligand interactions [44, 99], phosphorylation [100, 101], ubiquitination [102]), relies on decoding the information embedded within amino acid sequences and the rules governing their 3D structural formation [25, 44, 103].

This challenge aligns closely with AI research on information theory and representation learning. Given LLMs' strong sequence representation capabilities, the analogy between protein primary structures and natural language sequences has inspired researchers to model protein sequences within the LLM framework (pre-training phase) [10, 17]. Subsequently, these foundational models are fine-tuned for diverse protein-specific prediction and biological tasks (fine-tuning phase). Furthermore, integrating multimodal data, such as hierarchical spatial structures (from secondary to quaternary) [14, 38, 104] and functional annotations [105, 106], into LLMs has become a priority in recent research. Core focus areas in this field include the following.

(1) Development of optimized pre-trained large models: balancing performance with resource efficiency to handle diverse protein datasets and types.

(2) Effective fine-tuning methodologies: enabling targeted adaptations of LLMs for various protein-specific tasks.

Table 1 Summary of datasets in the genomics and transcriptomics field. Seq: DNA or RNA sequence; Func: functional annotations; Struc: structure; Mut: mutation (including SNP, insertions, and deletions).

Database	Data type	Scale	Cross-multiple species	Feature	Model fit example
Pre-training					
GRCh38 [78]	DNA Seq, gene annotations	Full human genome, 3.2G nucleotides [33]	Human	Genome reference	DNABERT, Enformer, DNABERT-2, DNAGPT
NCBI-Genome [70]	Biological Seq, Omics data (DNA, RNA, RNA-seq), Func annotation, etc.	Contain vast amounts of BiolocalSeq, Omics data cross-species	Yes	Organized, structured, various biological data	ENBED
Ensembl [79]	Biological Seq, Omics data (DNA, RNA, RNA-seq), Func annotation, etc.	Contain large amounts of BiolocalSeq, Omics data cross over 300 species, with a strong focus on vertebrates (less than NCBI-Genome)	Yes	Organized, structured, various biological data	SA DNA-LM [80]
1000 Genomes Project [81]	DNA Seq, Mut annotation	20.5T nucleotides [33], over 88 million SNPs and 1.4 million short insertions and deletions	Human	Human genetic variation across populations worldwide	Nucleotide-Transformer [82]
CGGA [83]	DNA Seq, RNA-seq, DNA methylation data	∼2k primary and recurrent glioma samples	Human	Genomic data focused on glioma patients from a Chinese cohort	–
ENCODE [84]	Biological Seq (DNA, RNA), Chromatin accessibility data, RNA-seq Annotation (Interaction, Func)	Encompasses over ∼14k types of experimental data from various tissues or cell lines, covering a wide array of sequencing data (such as RNA-seq)	Yes	Gene function and expression datasets	EpiGePT, GROVER
RNAcmap [28]	RNA contact map (Struc)	–	Yes	Automatic evolutionary coupling analysis for RNA sequence	RNA-MSM
BV-BRC [85]	Biological Seq, Omics data, annotation (Func, drug resistance, etc.)	Over 600k bacterial genomes, 1000 archaeal genomes, 8.5 million viral genomes	Yes	Bacterial and ViralPathogens, SARS-CoV-2 genomes	GenSLMs [94]
Panglao [86]	scRNA-seq	4M cells [33]	Human and mouse	scRNA-seq data	scBERT
Fine-tuning					
NT-Bench [82]	Genomic data (DNA), Annotation (TFBS, Promoters, and Enhancers site)	3202 diverse human genomes, 850 genomes from various species	Yes	Benchmark for evaluating the Transformer-based model in the DNA Seq task	DNABERT, Enformer, ENBED, Nucleotide-Transformer
PGB [73] (proposed)	Genomic data (DNA)	48 Plants genomics	Yes (plant species)	Benchmark for plant genomic research modeling.	AgroNT [73]
EPDnew [87]	DNA Promoter Seq	187k promoters [33]	Yes	Benchmark for Promoter seq prediction	DNABERT, miProBERT
iDNA-MS [88]	DNA methylation data	–	–	Supply benchmark dataset for 5hmC, 6mA, 4mC modification	BERT6MA, iDNA-ABT, MuLan-Methyl
CAGI5 [89]	Genomic data (DNA), Annotation (Func, single-nucleotide variants (SNV), etc.)	Over 810M potential nonsynonymous variants compared to reference genomics	Yes	Benchmark for evaluating effects on regulatory elements, prediction for genetic and genomic outcomes	–
ENCODE (protein-RNA binding) [90]	–	–	Yes	Include benchmark for RNA-binding protein (RBP) interactions prediction, (it is part of ENCODE databases)	BERT-RBP [91]
Zheng68k [92]	scRNA-seq, cell type annotation	∼6.8k cells	Human	Benchmark for cell type prediction	scBERT
STOmics DB [93]	Spatial transcriptomics data	228 spatial transcriptomic datasets	Yes	Storage and integration of spatial transcriptomic datasets	–

(3) Incorporation of evolutionary information: designing LLMs to represent and leverage protein evolutionary data, advancing research on protein evolution.

(4) Expanding applications in synthetic biology: utilizing LLMs to support protein design and engineering in synthetic biology.

4.2 Algorithm and application

4.2.1 *Pre-training algorithms*

Developing optimized pre-trained large models that balance performance and resource utilization is a significant challenge in protein modeling. Recent studies highlight that the key lies in enabling models to effectively handle diverse protein data types. Rather than solely fitting protein sequence data, integrating cross-domain information (e.g., protein annotations) and cross-modal information during pre-training has been shown to significantly enhance performance. By incorporating both protein sequences and gene ontology (GO) annotations [107] as input during pre-training, ProteinBERT [108] demonstrates improved performance across several downstream tasks. Importantly, this strategy achieves these gains without expanding model parameters, outperforming sequence-only models such as ESM-1b [109], ProtT5 [110], TAPE-Transformer [108], and UDSMProt [111] (a pre-trained LSTM model). ProtST [106] leverages a combination of protein sequences and biomedical text, adopting three masked pre-training techniques (unimodal and multimodal) to improve multimodal information alignment. This allows ProtST to excel in retrieving functional proteins from large databases. ProGen [112] and ProGen2 [113] integrate protein sequences with biological domain conditioning tags (e.g., taxonomic and keyword labels) during pre-training to enhance protein design capabilities. ProGen2 builds upon ProGen by expanding its parameter count and training data, achieving further improvements in protein design tasks.

Additionally, non-sequential modality information can also be generated from the sequence itself. Non-sequential information, such as protein 3D structure, can also be derived directly from sequences and incorporated into pre-training. Using a structure-aware (SA) approach, SaProt [37] employs Foldseek [114] to generate 3D structural information tokens. These are integrated with protein sequence data, enabling SaProt to outperform ESM2 in tasks like protein contact prediction. Additionally, SaProt demonstrates superior zero-shot mutational effect inference, particularly when compared to ESM2, whose performance does not scale proportionally with size increases (e.g., from 650 million to 3 billion parameters). Building on ProtT5, ProtST5 [38] incorporates structural tokens generated by Foldseek and is trained on billions of protein sequences using span corruption techniques. This model surpasses AlphaFold and ESMFold (fine-tuned from ESM2) in structural generation tasks, demonstrating its efficiency in capturing both sequence and structural information.

In contrast, large foundational models (including single-modality models) can also enhance performance by focusing on extracting relational information between protein sequences, emphasizing sequence attribution. A prominent example is the AlphaFold family, which relies on a core architectural design centered around relational modeling between sequences. PoET [15] employs a unique "sequence-of-sequences" approach, concatenating multiple sequences from the same protein family to capture hierarchical evolutionary relationships. With its intra-sequence and inter-sequence module designs, PoET can exceed the sequence lengths encountered during pre-training. PoET surpasses models like ESM-1v [109] in evolutionary prediction tasks (e.g., mutation effects) and demonstrates strong performance in functional sequence design tasks.

While integrating cross-domain and multimodal information can enhance model performance, the scarcity of multimodal and annotated data poses a significant challenge. Models such as ESM3 and SaProt, which focus on generating 3D structural representations from sequences, circumvent this issue by relying primarily on sequence-derived information. These methods are particularly valuable in scenarios where high-quality non-sequential data are unavailable. However, recent research has raised concerns about the risks associated with "generated" data in LLMs [115]. Addressing the scarcity of cross-modal data and understanding the extent to which sequence-derived information can represent other modalities remain critical challenges for developing more robust and versatile LLMs for protein modeling.

4.2.2 *Fine-tuning algorithms*

The success of fine-tuning LLMs for domain-specific tasks depends on selecting appropriate methodologies that enable effective task-specific data embedding and representation. Protein-related downstream tasks

can generally be divided into sequence-level tasks (e.g., subcellular localization prediction [22, 23]) and amino acid residue (token)-level tasks (e.g., protein-protein interaction sites [24], ubiquitination modification sites [116], or protein structural predictions [7, 14, 17, 44, 117]). Each type of task often requires tailored fine-tuning methods. Fine-tuning methodologies can be categorized into two main approaches.

(1) Feature extraction, where representations from hidden layers of the LLM are used as inputs to separate downstream models.

(2) Direct weight fine-tuning, where the model's parameters are updated directly using downstream data, including approaches like full fine-tuning and parameter-efficient fine-tuning.

For feature extraction methods addressing sequence-level problems, it is necessary to construct an overall representation of the sequence. The most widely used approach involves global average pooling based on features from all amino acid residues. For instance, Elnaggar et al. [7] extracted amino acid features from ProtT5's last hidden layer, applied global average pooling, and fed the resulting sequence representation into a DNN to predict subcellular localization. Fang et al. [118] used global average pooled features from ProtT5 in their AFP-MFL model to identify antifungal peptides with a simple MLP. Wang et al. [119] compared features extracted from antiBERTy [120], ProtT5, and ESM2 for B-cell receptor (BCR) sequences and demonstrated that task-specific fine-tuning led to better feature representations. ProtLoc-Mex [22] introduced a novel approach by combining special character embeddings and segmental average pooling of features extracted from ESM2 for subcellular localization prediction. Additionally, overall sequence representation features can also be aggregated through attention mechanisms. DeepLoc 2.0 [23] extracted residue-level features from ProtT5 and ESM2 (650M) and aggregated them using self-attention to build a subcellular localization model. MFE [121] used cross-attention to integrate sequence features (from ProtBERT) with molecular surface point cloud features for protein-ligand binding affinity prediction.

For amino acid residue-level tasks, fine-tuning typically avoids aggregating residue features into a single representation. Instead, task-specific models are used to process relationships between residue features directly. Elnaggar et al. [7] extracted amino acid features from ProtT5 and employed a CNN to predict per-residue secondary structure labels. DeepProSite [24] extracts protein sequence features from ProtT5 and structural features of proteins from ESMFold [17], then integrates these cross-modal features by a graphic Transformer to predict protein binding sites. In contrast, when directly fitting LLMs to downstream sequence-level and residue-level tasks, it is crucial to first determine whether the task involves new sequence or residue (token) generation. Generally, sequence-level tasks do not require generative modeling. Instead, these tasks leverage features extracted from special tokens like the CLS (classification) token or global average pooling. These features are then processed by a multi-layer perceptron (MLP) for prediction. For residue-level tasks, the approach varies depending on whether generative modeling is involved. Non-generative tasks (e.g., protein binding site prediction) resemble named entity recognition (NER) tasks in NLP, where models like BERT process residue-level features sequentially through an MLP to assign labels. Conversely, generative residue-level tasks require training with a generative head or decoder. In such cases, GPT models, owing to their generative pre-training paradigm, require minimal architectural modifications during fine-tuning to handle these tasks.

In summary, feature extraction-based methods have demonstrated success across various tasks, as features can be fine-tuned independently of the larger model. Designing features based on specific attributes or integrating multimodal features (e.g., MFE, DeepProSite, and EasIFA [122]) can further enhance downstream model performance. Additionally, selecting the appropriate downstream model is critical. Refs. [10, 22, 25, 105] have shown that simpler models, such as support vector machines (SVMs), random forests (RFs), or logistic regression classifiers, can offer competitive predictive performance while effectively reflecting the properties of feature representations. Furthermore, the characteristics of the fine-tuning dataset, including class balance, significantly influence downstream model performance. Besides, directly fine-tuning the LLM parameters for downstream tasks can achieve superior performance, but often requires robust open-source support for the models. Direct fine-tuning involves backpropagation, which is computationally intensive and demands substantial GPU memory, making it less accessible to researchers not involved in the pre-training phase. Although feature extraction methods are predominantly used with the ESM-family and ProtTrans-family, solutions such as PEFT techniques and model distillation methods, like DistilProtBERT [123], offer promising alternatives for researchers facing resource constraints.

4.2.3 *Evolutionary information representation algorithms*

Proteins are not merely static sequences; they carry rich evolutionary information, which is essential for improving model performance and deepening our understanding of protein functionality. Capturing this evolutionary information has become a key focus in protein representation learning (PRL) [10, 16, 25, 116], including applications leveraging LLMs. Classical methods, such as multiple sequence alignments (MSA) [124], position-specific scoring matrices (PSSM) [125], and BLOSUM62 [118], have been widely used to represent protein features that reflect evolutionary information. These techniques effectively capture conserved regions and mutation patterns across protein families, enabling downstream tasks like function prediction, structure determination, and protein design. Recent advancements have incorporated the MSA module directly into deep learning architectures to enhance protein representation learning. For instance, AlphaFold leverages MSA-based representations to significantly improve protein structure prediction accuracy. EvoDiff [126] integrates MSA modules to capture evolutionary relationships, demonstrating the utility of explicit evolutionary information in model performance.

While LLMs were originally designed to generate general protein representations, certain architectural modifications enable explicit incorporation of evolutionary information, even without relying on MSA matrices. PoET [15] introduces sequence concatenation from protein families as input, enabling the model to explicitly capture hierarchical evolutionary relationships. This approach infuses evolutionary information into the model through task-specific design, bridging the gap between LLMs and traditional alignment-based methods. Interestingly, research has also revealed that LLMs without explicit evolutionary modules can implicitly capture the evolutionary properties of proteins. This phenomenon likely stems from two key factors.

(1) Diverse pre-training data: LLMs are trained on extensive protein datasets spanning broad evolutionary distances, allowing them to learn patterns that inherently reflect evolutionary relationships [14, 17].

(2) Self-supervised learning objectives: Tasks like reconstructing masked residues or predicting subsequent residues during pre-training allow LLMs to infer reconstruction rules. These rules often parallel the natural evolutionary processes that shape proteins [14].

It is noteworthy that in the field of NLP, LLMs have exhibited emergent capabilities when their parameter scale surpasses a certain threshold—an observation known as "emergence" [127] and a key aspect of scaling laws theory. Similarly, the expansion of ESM3 to 98 billion parameters has significantly enhanced protein representation performance and improved accuracy in downstream tasks [14]. Notably, ESM3 has demonstrated robust generative capabilities across evolutionary distances, suggesting that scaling laws may give rise to "emergent" abilities in extracting evolutionary information from protein sequences. This insight, based on the 7-billion-parameter version used in recent experiments, holds considerable implications for advancing protein-focused large language models. Such findings could guide future algorithmic research aimed at encoding evolutionary information within models and improving protein design by mirroring evolutionary principles. The unique GFP protein generated by ESM3 exemplifies this potential, as it represents a remarkable achievement in life sciences research and spans an evolutionary distance of over 500 million years [128].

While MSA-based methods remain foundational in evolutionary information modeling, LLMs offer a complementary pathway by learning evolutionary relationships directly from large-scale data. Future research may focus on the following.

- Developing hybrid approaches that integrate explicit alignment-based evolutionary modules with LLMs' implicit learning capabilities.
- Scaling parameter counts in LLMs to explore emergent properties further.
- Designing tasks and architectures that better reflect evolutionary rules, enabling LLMs to generate novel proteins with precise evolutionary context.

4.2.4 *Application in synthetic biology*

LLMs are emerging as transformative tools in synthetic biology research. In this field, three progressively advanced tasks highlight their potential: decoding and annotating protein functions, high-precision structure prediction, and functional-oriented protein design. First, LLMs excel in predicting protein functions and generating high-quality annotations [129]. By leveraging extensive pre-trained knowledge, they enable large-scale and accurate functional analysis of proteins. Second, LLMs have proven adept at capturing the rules governing protein structure prediction [7, 14, 17, 112, 113]. Models such as AlphaFold and ESMFold have redefined structural biology by accurately predicting protein 3D conformations. Finally, LLMs

demonstrate remarkable capabilities in functional-oriented protein design [14, 15, 103, 130, 131]. This includes designing functional scaffolds, engineering structural motifs tailored for specific applications [15], and generating entirely novel proteins not found in nature [17]. With their breakthrough abilities, LLMs are becoming indispensable tools for realizing the long-sought vision of programmable protein design [132], revolutionizing synthetic biology, and expanding its horizons.

4.2.5 *Application in evolutionary biology*

Integrating evolutionary information into protein representation models has the potential to profoundly advance the field of protein research by enriching model understanding of sequence-function relationships. Evolutionary signals, such as residue conservation and co-evolution, provide critical context for identifying functionally important regions, predicting the effects of mutations, and inferring structural and interaction features. These enriched models can lead to more accurate annotation of uncharacterized proteins, improved variant interpretation in clinical genomics, and enhanced capabilities in rational protein engineering and drug design. Additionally, leveraging evolutionary data across species supports comparative proteomics and the reconstruction of ancestral proteins, deepening our understanding of protein evolution and diversity.

In the field of protein evolutionary biology, LLMs are becoming a research hotspot [14, 26, 132]. These models excel at embedding evolutionary information and providing high-quality representations of protein sequences. As a result, fine-tuned LLMs are highly effective in accomplishing tasks related to protein evolution analysis, such as identifying conserved structural domains [133] and conducting homology comparisons [134]. For instance, Yeung et al. [133] utilized LLMs like ProtT5, ESM-1b, and ESM2 to extract protein feature representations, which were then input into downstream prediction models to estimate protein sequence conservation domains. Similarly, CATHe leverages feature representations extracted from ProtT5 to detect distant homologs within the CATH superfamily. PLMSearch [26], a protein language model-based framework, achieves high speed and sensitivity in homologous protein searches while excelling at detecting distant homologs. A significant emerging trend involves calculating the similarity of protein feature representations in vector space for evolutionary modeling, rather than relying on traditional sequence alignment methods. This approach exemplifies a bioinformatics revolution powered by LLMs, enabling more efficient and scalable evolutionary analyses.

4.3 Database

Most protein datasets primarily store amino acid sequences (protein primary structure), while some protein databases also include multimodal data such as 3D spatial structures. Table 2 [27,41,133,135–155] delineates commonly employed datasets for pretraining or fine-tuning LLMs in the domain of proteomics, highlighting the datasets' attributes, application, and scenarios.

4.4 Summary

This section delves into the utilization of LLMs in proteomics, highlighting their purpose, various model pre-training and fine-tuning methods, and the differences in algorithms rooted in evolutionary information representation. The cutting-edge applications of LLMs in proteomics are also examined, alongside a summary of relevant datasets. As the latest advancement in the protein language model (PLM) [7] family, LLMs have significantly improved the modeling of protein sequences and the integration of multimodal data in proteomics. These models, however, are now encountering new challenges from emerging architectures, such as diffusion models [126]. Therefore, future efforts may need to focus on enhancing LLM architectures and exploring integration with other models to further advance their capabilities [44, 156].

5 Challenges in current biological LLMs and future directions

While significant breakthroughs have been achieved with LLMs in genomics, transcriptomics, and proteomics, several challenges remain. These include issues related to interpretability, causality, data quality, computational resources, hallucinations, and security.

Table 2 Summary of datasets in the proteomics field. Seq: protein sequence; Fam: protein family; Dom: protein domain; Cluster: protein cluster information; Func: functional annotations; Struc: protein structure; IHC: immunohistochemistry.

Database	Data type	Scale (protein)	Cross-multiple species	Feature	Model fit example
Pre-training					
UniRef100 [135, 136]	Seq, Cluster, Func annotation	412M	Yes	Cluster identical sequences and fragments	ProtTrans-family
UniRef90 [135, 136]	Seq, Cluster, Func annotation	192M	Yes	Cluster sequences with 90% identity, 80% overlap within the UniRef100 database	ESM-family
UniRef50 [135, 136]	Seq, Cluster, Func annotation	66M	Yes	Cluster sequences with 50% identity, 80% overlap within the UniRef100 database	ESM-family, POET
Pfam [137]	Seq, Fam, Dom info	50M	Yes	Protein family datasets, MSA-based search	PLMSearch
Swiss-Prot (UniProtKB) [138]	Seq, annotation (Loc, PPI, Go, etc.)	570k	Yes	Database for protein storage, manually curated with high-quality protein sequences and annotations	ProGen, DeepLoc 2.0
TrEMBL (UniProtKB) [138]	Seq, Annotation (Loc, PPI, Go, etc.)	250M	Yes	Database for protein storage, computationally annotated in protein sequence	–
PDB [27]	3D structure, annotation	214k	Yes	High-resolution 3D structures from X-ray crystallography, NMR, and cryo-EM	AlphaFold-family, ESM-family, ProstT5
AlphaFoldDB [41, 139]	Predict Struc	214M	Yes	Provides high-confidence structural predictions, but not experimentally determined	AlphaFold-family, ESM-family
BFD [140, 141]	Seq, Cluster	2.5B	Yes	Storing large and comprehensive sequences sourced from multiple databases	ProtT5, ProtBERT
UniParc [142]	Seq	–	Yes	Non-redundant and large protein sequence database	ProGen, TransPTM [133]
Fine-tuning					
GO [143]	Annotation (BP, CC, MF)	1.5M	Yes	Dynamic, hierarchically structured biological ontology knowledge base	Protein BERT
EC [144]	Enzyme classification annotation	2.6M	Yes	Provides a systematic way of naming and categorizing enzymes based on the reactions they catalyze	PLMs-based Framework (Guisheng Fan, etc.) [145]
HPA [146]	Annotation (Loc, etc.), metadata (IHC, RNA-seq)	–	Human	Benchmarks for evaluating sub-localization prediction	DeepLoc 2.0
STRING [147]	Seq, annotation (PPI, etc.)	59M	Yes	Benchmarks for evaluating protein-protein interaction	ProLLM [148]
CATH [149]	Dom and Fam annotation	151M	Yes	Hierarchical classification based on structure and evolutionary relationships	ProstT5
TAPE [150]	Secondary Struc,annotation (contact, landscape, remote homology)	120k	Yes	Benchmarks for evaluating protein representation	PTG-PLM [151]
ProteinGym [152]	Deep mutational scanning (DMS) assays	300k	Yes	Benchmark for evaluating protein design and fitness, predicts how mutations affect protein stability and function	Saprot
FLIP [153]	Annotation (Adenovirus Stability, Stability of Protein Domain B1, Fam, etc.)	320k	Yes	Benchmarks for evaluating protein landscape prediction	–
PLMD [154]	Annotation: post-translational modifications (PTMs) of lysine residues	121k ubiquitination sites across 25k protein Seq	Yes	Datasets for PTMs prediction	–
CPLM 4.0 [155]	Annotation: post-translational modifications (PTMs) of lysine residues	592k modification events, 463k unique lysine residues, 105k protein Seq	Yes	Benchmarks for PTMs prediction	–

5.1 Data quality and computational resources

High-quality data are essential for training large models, but obtaining such datasets in biological omics is challenging due to issues like batch effects in RNA-seq data [29] and the impact of sequence mutations on tokenization [72]. There are also concerns about the potential for bias in training data. Biological datasets, especially those derived from public repositories, often reflect skewed sampling across species, tissues, or disease types, which may lead LLMs to learn and propagate these biases. This could result in inaccurate predictions or reduced generalizability to underrepresented biological contexts. As LLMs are increasingly applied in clinical and biomedical settings, mitigating such biases through data augmentation, balanced sampling, or fairness-aware learning techniques will be crucial.

Additionally, pre-training LLMs is resource-intensive, and scaling large models is often beyond the reach of resource-limited research groups, particularly in the biological domain. For example, most biological LLMs have fewer than 10 billion parameters, with the largest models not exceeding 100 billion parameters, while general LLMs such as Llama3 have up to 405 billion parameters. Fine-tuning larger models is also difficult due to higher-dimensional feature vectors and the vast number of tokens in biological sequences, which create extensive feature matrices that burden LLM inference and fine-tuning processes. Developing more efficient architectures, parameter-efficient fine-tuning strategies (such as LoRA or adapters), and model distillation techniques may help democratize access to powerful biological LLMs while reducing resource consumption.

Future direction. To address the scarcity of high-quality data, approaches like zero-shot and few-shot learning [157,158] may be useful. Additionally, to mitigate resource consumption during pre-training and fine-tuning, methods such as deploying variational autoencoders (VAE) to reduce dimensionality [22] or employing distillation techniques [123] to train high-accuracy smaller models could serve as alternatives to large models.

5.2 Hallucinations and security

Generative models like LLMs and diffusion models face challenges related to "hallucination" issues [44, 126]. These issues not only affect the quality of generated outputs but also pose security risks, as they can result in incorrect functional labels for proteins, leading to potentially misleading or harmful biological research outcomes. Furthermore, these models can also amplify biases [159], raise privacy concerns, and present other ethical dilemmas. For example, protein drugs generated with the aid of these models may exhibit differential effectiveness across ethnic groups if the training data are skewed towards specific demographic sensitivities [31,160]. There is also the potential risk of sensitive genomic data leakage [161]. Moreover, these models could facilitate the development of biochemical weapons by enhancing the analysis and prediction of synthetic functional structures.

Future direction. To address these issues, potential solutions may include aligning models with techniques such as retrieval-augmented generation (RAG) [15] and reinforcement learning with human feedback (RLHF) [1] to improve model reliability and safety.

5.3 Interpretability and causality

LLMs have the potential to "replace" intricate and often unknowable gene expression regulation processes, thus facilitating a range of downstream tasks, such as predicting phenotypes (illustrated in Figure 4). However, biological research seeks not only to construct predictive models but also to understand the complex mechanisms of gene regulation [162]. It aims to explore how models "comprehend" biological data [22, 99]. Despite their utility, LLMs often lack interpretability and cannot directly capture causal relationships within the data, rendering them "black boxes" (illustrated in Figure 4).

Future direction. To enhance interpretability, one approach involves developing self-explanatory LLMs, also known as "white-box" models [163]. Another approach could involve extracting feature semantics using sparse autoencoders (SAE) [3], and applying feature attribution methods like SHAP [164], LIME [165], and attention mechanisms, which may help in interpreting LLMs [22]. However, feature attribution methods can face challenges in robustness and were not initially designed for LLMs interpretation. For integrating causal inference into LLMs, a potential solution could involve incorporating causal knowledge of gene expression regulation into structured causal language data and then fitting LLMs to these data using a chain-of-thought (COT) inference framework [166]. This emerging field, known as causal machine learning, holds promise for future advancements in this area [167].

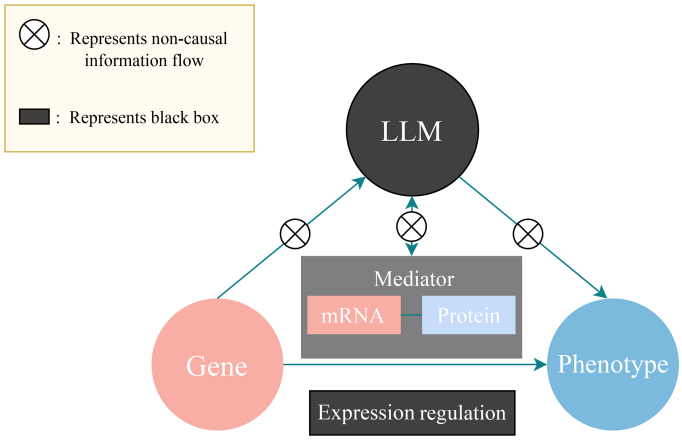

Figure 4 (Color online) LLMs as a tool to elucidate the biological process linking genes to phenotypes. The black box indicates that both the gene expression regulation processes and the predictive decision-making mechanisms of LLMs are too complex to allow for complete or reasonable explanations. The gray box (mediator), represented by RNA and protein, signifies that biological experiments can leverage these intermediate processes in the pathway from genes to phenotypes. Research on RNA and proteins can partially address the black-box nature of phenotype regulation. This concept could inform future LLM advancements, emphasizing the integration of DNA, RNA, and protein data to enhance understanding and predictive accuracy in biological research.

5.4 Future directions in LLM integration

A promising, yet developing application is the integration of general-purpose LLMs such as ChatGPT [1], Claude [3], and Gemini [168] within the biological omics research space. This is demonstrated by the development of single-cell annotation protocols based on the OpenAI GPT-4 API [169], conducting biological knowledge retrieval via natural language [170], and using GPT-4 to generate protein structures using prompts alone [171]. The potential of model APIs or natural language interaction via prompt engineering [172], which enables general LLMs to be transferred into new bioinformatics and omics research areas without fine-tuning, is a highly promising direction.

Moreover, integrating more multimodal and interdisciplinary data related to biological phenotypes, such as data from animals [173], plants [64], and clinical samples [160, 174, 175], could enable LLMs to better address challenges in genomics, transcriptomics, and proteomics. This potential is supported by the vast information compression capabilities of LLMs and the demonstrated versatility of existing general models, such as GPT-4, trained across multiple domains. Since biological organism phenotypes represent the final expression of genomic, transcriptomic, and proteomic data, one question that arises is whether increasing the data volume will lead to greater knowledge and result in more powerful models.

Specifically, some research has explored the combination of biological omics data with clinical and pathological data for disease diagnosis and treatment [176–179]. Among these, image data (e.g., whole-slide images, or WSI) play a critical role in the clinical domain. As Transformers are capable of learning from visual sequences, audio sequences, and other data represented in sequential form [180], recent advancements have focused on segmenting images into patches, which are then treated as tokens to construct sequences [181, 182]. These sequences are modeled by Transformers, leveraging a pretraining-finetuning approach to create unified multimodal models that integrate imaging and sequential data [183] (known as computational pathology LLMs). These models have already been successfully applied to clinical downstream prediction tasks. Moving forward, integrating biological sequence data and RNA-seq data with image data to develop foundational computational pathology models holds great promise. Such models could unify information representation through contrastive learning across multimodal data and disciplines.

Finally, the open-source availability of LLMs is crucial for the sustained integration of these large-scale models into biological research. A key example in the open-source community is the ESM family, including ESM-1b, ESM2, ESM3, and fine-tuned models such as ESM-IF1, ESM-1v, and ESMfold. These models play a critical role in the open-source ecosystem of proteomic LLMs and facilitate model utilization and redevelopment across a broad spectrum.

6 Conclusion

This review thoroughly explores the mission, architecture, datasets, and advanced applications of deploying LLMs in the fields of genomics, transcriptomics, and proteomics. We provide detailed introductions to the basic architecture of LLMs and discuss the modifications required to adapt these models for biological applications, focusing on how they handle complex sequence-based and non-sequence-based domain data. We also examine the applications of LLMs in downstream tasks across these fields, highlighting the technical nuances that arise when using these models in biological research. Furthermore, the significance of transfer learning and the techniques behind transfer methods are essential to understanding the success of LLMs in biological omics tasks. Finally, we address several challenges and future directions for the use of LLMs in biological omics research. These challenges include issues related to data quality, interpretability, and computational resources, as well as concerns about model hallucinations and security. Despite these challenges, the future of LLMs in this domain holds immense potential, particularly with the integration of multimodal data and the scaling of models to better capture the complexities of biological systems.

Humanity has long sought to decode the vast information encoded in the living world, from the simplest "progenitor cells" to the diverse species present today. Just as humans have created intricate systems for storing and communicating information, such as language and imagery, the advent of LLMs in AI marks a significant leap forward in our ability to understand and utilize this information. As AI continues to advance, it may enable machines to better decode the mysteries of life's information, enhancing both our understanding of biology and the technologies we use to study it. This ongoing evolution in AI research represents a promising path toward further unraveling the complexities of life and our place within it.

Acknowledgements We acknowledge the contributions of the open-source Protein Quaternary Structure and Protein Tertiary Structure icons, developed by the DBCLS authors (https://togotv.dbcls.jp/en/pics.html), which are available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Supporting information Table A1. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. OpenAI, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
2 Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinf, 2024, 25: bbad493
3 Templeton A, Conerly T, Marcus J, et al. Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. Transform Circuits Thread, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
4 Vaswani A, Shazeer N M, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 6000–6010
5 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, 2019. 4171–4186
6 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 5485–5551
7 Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 7112–7127
8 Manzoni C, Kia D A, Vandrovcova J, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinf, 2018, 19: 286–302
9 Wu L, Huang Y, Lin H, et al. A survey on protein representation learning: retrospect and prospect. 2022. ArXiv:2301.00813
10 Detlefsen N S, Hauberg S, Boomsma W. Learning meaningful representations of protein sequences. Nat Commun, 2022, 13: 1914
11 Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods, 2021, 18: 1196–1203
12 Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics, 2021, 37: 2112–2120
13 Sanabria M, Hirsch J, Joubert P M, et al. DNA language model GROVER learns sequence context in the human genome. Nat Mach Intell, 2024, 6: 911–923
14 Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. Science, 2025, 387: 850–858
15 Truong T F, Bepler T. PoET: a generative model of protein families as sequences-of-sequences. 2023. ArXiv:2306.06156
16 Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA, 2021, 118: e2016239118
17 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379: 1123–1130
18 Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 2024, 21: 1470–1480
19 Corces M R, Granja J M, Shams S, et al. The chromatin accessibility landscape of primary human cancers. Science, 2018, 362: eaav1898
20 Gao Z, Liu Q, Zeng W, et al. EpiGePT: a pretrained transformer model for epigenomics. bioRxiv, 2023. doi: 10.1101/2023.07.15.549134
21 Jin J, Yu Y, Wang R, et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. Genome Biol, 2022, 23: 219

22 Luo Z, Wang R, Sun Y, et al. Interpretable feature extraction and dimensionality reduction in ESM2 for protein localization prediction. Brief Bioinf, 2024, 25: bbad534
23 Thumuluri V, Armenteros J J A, Johansen A R, et al. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res, 2022, 50: W228–W234
24 Fang Y, Jiang Y, Wei L, et al. DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. Bioinformatics, 2023, 39: btad718
25 Wang M, Patsenker J, Li H, et al. Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. Nucleic Acids Res, 2024, 52: 548–557
26 Liu W, Wang Z, You R, et al. PLMSearch: protein language model powers accurate and fast sequence search for remote homology. Nat Commun, 2024, 15: 2775
27 Burley S K, Berman H M, Bhikadiya C, et al. Protein data bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res, 2019, 47: D520–D528
28 Zhang T, Singh J, Litfin T, et al. RNAcmap: a fully automatic pipeline for predicting contact maps of RNAs by evolutionary coupling analysis. Bioinformatics, 2021, 37: 3494–3500
29 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods, 2024, 21: 1481–1491
30 Yang F, Wang W, Wang F, 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: 852–866
31 Wang B, Xie Q, Pei J, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv, 2023, 56: 1–52
32 Liu J, Yang M, Yu Y, et al. Large language models in bioinformatics: applications and perspectives. 2024. ArXiv:2401.05632
33 Zhang Q, Ding K, Lyv T, et al. Scientific large language models: a survey on biological & chemical domains. 2024. ArXiv:2401.14656
34 Zhang Y, Chen X, Jin B, et al. A comprehensive survey of scientific large language models and their applications in scientific discovery. 2024. ArXiv:2406.10833
35 Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022. 35: 1950–1965
36 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022. 35: 27730–27744
37 Su J, Han C, Zhou Y, et al. SaProt: protein language modeling with structure-aware vocabulary. bioRxiv, 2023. doi: 10.1101/2023.10.01.560349
38 Heinzinger M, Weissenow K, Sanchez J G, et al. Bilingual language model for protein sequence and structure. NAR Genomics Bioinf, 2024, 6: lqae150
39 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. 2015, ArXiv:1508.07909
40 Choromanski K, Likhosherstov V, Dohan D, et al. Rethinking attention with performers. 2020. ArXiv:2009.14794
41 Varadi M, Bertoni D, Magana P, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res, 2024, 52: D368–D375
42 Su J, Ahmed M, Lu Y, et al. RoFormer: enhanced transformer with rotary position embedding. Neurocomputing, 2024, 568: 127063
43 Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized Bert pretraining approach. 2019. ArXiv:1907.11692
44 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493–500
45 van der Auwera G A, O'Connor B D. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. Sebastopol: O'Reilly Media, 2020. 1–500
46 Zhu P, Shu H, Wang Y, et al. MAEST: accurately spatial domain detection in spatial transcriptomics with graph masked autoencoder. Brief Bioinf, 2025, 26: bbaf086
47 Zhang T, Zhang X, Wu Z, et al. VGAE-CCI: variational graph autoencoder-based construction of 3D spatial cell-cell communication network. Brief Bioinf, 2025, 26: bbae619
48 Yang C, Liu Y, Wang X, et al. stSNV: a comprehensive resource of SNVs in spatial transcriptome. Nucleic Acids Res, 2025, 53: D1224–D1234
49 Wang T, Shu H, Hu J, et al. Accurately deciphering spatial domains for spatially resolved transcriptomics with stCluster. Brief Bioinf, 2024, 25: bbae329
50 Wang T, Mai D, Shu H, et al. Enhancing cell subpopulation discovery in cancer by integrating single-cell transcriptome and expressed variants. Fundamental Res, 2025, 5: 1–10
51 Wang T, Zhao H, Xu Y, et al. scMultiGAN: cell-specific imputation for single-cell transcriptomes with multiple deep generative adversarial networks. Brief Bioinf, 2023, 24: bbad384
52 Zhang D, Zhang W, He B, et al. DNAGPT: a generalized pretrained tool for multiple DNA sequence analysis tasks. bioRxiv, 2023. doi: 10.1101/2023.07.11.548628
53 Song X, Salcianu A, Song Y, et al. Fast wordpiece tokenization. 2020. ArXiv:2012.15524
54 Bostrom K, Durrett G. Byte pair encoding is suboptimal for language model pretraining. 2020. ArXiv:2004.03720
55 Zhou Z, Ji Y, Li W, et al. DNABERT-2: efficient foundation model and benchmark for multi-species genome. 2023. ArXiv:2306.15006
56 Cao C, Wang C, Dai Q, et al. CRBPSA: CircRNA-RBP interaction sites identification using sequence structural attention model. BMC Biol, 2024, 22: 260
57 Saha B, Ye C. The I/O complexity of attention, or how optimal is flash attention? 2024. ArXiv:2402.07443
58 Romero I G, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet, 2012, 13: 505–516
59 Costello J C, Heiser L M, Georgii E, et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol, 2014, 32: 1202–1212
60 Wang T, Yang J, Xiao Y, et al. DFinder: a novel end-to-end graph embedding-based method to identify drug-food interactions. Bioinformatics, 2023, 39: btac837
61 Leirisalo-Repo M. Prognosis, course of disease, and treatment of the spondyloarthropathies. Rheumatic Dis Clin North Am, 1998, 24: 737–751
62 Wang X, Gao X, Wang G, et al. miProBERT: identification of microRNA promoters based on the pre-trained model BERT. Brief Bioinf, 2023, 24: bbad093
63 Tsai M S, Lin M H, Lee C P, et al. Chang Gung Research Database: a multi-institutional database consisting of original medical records. Biomed J, 2017, 40: 263–269
64 Yoon Y S, Ahn H S, Allison P S, et al. Cosmic-ray proton and helium spectra from the first cream flight. Astrophys J, 2011, 728: 122
65 Tsukiyama S, Hasan M M, Deng H W, et al. BERT6mA: prediction of DNA N6-methyladenine site using deep learning-based approaches. Brief Bioinf, 2022, 23: bbac053

66 Yu Y, He W, Jin J, et al. iDNA-ABT: advanced deep learning model for detecting DNA methylation with adaptive features and transductive information maximization. Bioinformatics, 2021, 37: 4603–4610
67 Zeng W, Gautam A, Huson D H. MuLan-Methyl—multiple transformer-based language models for accurate DNA methylation prediction. Gigascience, 2023, 12: giad054
68 Clark K, Luong M T, Le Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators. 2020. ArXiv:2003.10555
69 Zhang W, Lazar-Stefanita L, Yamashita H, et al. Manipulating the 3D organization of the largest synthetic yeast chromosome. Mol Cell, 2023, 83: 4424–4437
70 Zhao Y, Coelho C, Hughes A L, et al. Debugging and consolidating multiple synthetic chromosomes reveals combinatorial genetic interactions. Cell, 2023, 186: 5220–5236
71 Schindler D, Walker R S K, Jiang S, et al. Design, construction, and functional characterization of a tRNA neochromosome in yeast. Cell, 2023, 186: 5237–5253
72 Malusare A, Kothandaraman H, Tamboli D, et al. Understanding the natural language of DNA using encoder-decoder foundation models with byte-level precision. 2023. ArXiv:2306.14999
73 Mendoza-Revilla J, Trop E, Gonzalez L, et al. A foundational large language model for edible plant genomes. Commun Biol, 2024, 7: 835
74 Yin W, Zhang Z, He L, et al. ERNIE-RNA: an RNA language model with structure-enhanced representations. bioRxiv, 2024. doi: 10.1101/2024.03.17.585376
75 Zhang Y, Lang M, Jiang J, et al. Multiple sequence alignment-based RNA language model and its application to structural inference. Nucleic Acids Res, 2024, 52: e3
76 Chen J, Hu Z, Sun S, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. 2022. ArXiv:2204.00300
77 Huang K, Qu Y, Cousins H, et al. Crispr-GPT: an LLM agent for automated design of gene-editing experiments. 2024. ArXiv:2404.18021
78 Guo Y, Dai Y, Yu H, et al. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics, 2017, 109: 83–90
79 Lamesch P, Berardini T Z, Li D, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res, 2012, 40: D1202–D1210
80 Gankin D, Karollus A, Grosshauser M, et al. Species-aware DNA language modeling. bioRxiv, 2023. doi: 10.1101/2023.01.26.525670
81 Auton A, Abecasis G R, Altshuler D M, et al. A global reference for human genetic variation. Nature, 2015, 526: 68–74
82 Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. Nat Methods, 2025, 22: 287–297
83 Zhao Z, Zhang K N, Wang Q, et al. Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese glioma patients. Genomics Proteomics Bioinf, 2021, 19: 1–12
84 Consortium E P. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74
85 Olson R D, Assaf R, Brettin T, et al. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic Acids Res, 2023, 51: D678–D689
86 Franzén O, Gan L M, Björkegren J L M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database, 2019, 2019: baz046
87 Dreos R, Ambrosini G, Périer R C, et al. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. Nucleic Acids Res, 2013, 41: D157–D164
88 Lv H, Dao F Y, Zhang D, et al. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. iScience, 2020, 23: 100991
89 Katsonis P, Lichtarge O. CAGI5: objective performance assessments of predictions based on the evolutionary action equation. Hum Mutat, 2019, 40: 1436–1454
90 Pan X, Fang Y, Li X, et al. RBPsuite: RNA-protein binding sites prediction suite based on deep learning. BMC Genomics, 2020, 21: 884
91 Yamada K, Hamada M, Arighi C. Prediction of RNA-protein interactions using a nucleotide language model. Bioinf Adv, 2022, 2: vbac023
92 Zheng G X Y, Terry J M, Belgrader P, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun, 2017, 8: 14049
93 Xu Z, Wang W, Yang T, et al. STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. Nucleic Acids Res, 2024, 52: D1053–D1061
94 Zvyagin M, Brace A, Hippe K, et al. GenSLMs: genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. Int J High Perform Comput Appl, 2023, 37: 683–705
95 Liu Y, Rosikiewicz W, Pan Z, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol, 2021, 22: 295
96 Zhao E, Stone M R, Ren X, et al. Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol, 2021, 39: 1375–1384
97 Zahedi R, Ghamsari R, Argha A, et al. Deep learning in spatially resolved transcriptomics: a comprehensive technical view. Brief Bioinf, 2024, 25: bbae082
98 Goodsell D S. The Machinery of Life. New York: Springer, 2009. 371–402
99 Mastropietro A, Pasculli G, Bajorath J. Learning characteristics of graph neural networks predicting protein-ligand affinities. Nat Mach Intell, 2023, 5: 1427–1436
100 Blom N, Sicheritz-Pontén T, Gupta R, et al. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 2004, 4: 1633–1649
101 Chandra A, Tünnermann L, Löfstedt T, et al. Transformer-based deep learning for predicting protein properties in the life sciences. eLife, 2023, 12: e82819
102 Fu H, Yang Y, Wang X, et al. DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins. BMC Bioinf, 2019, 20: 86
103 Gligorijević V, Renfrew P D, Kosciolek T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun, 2021, 12: 3168
104 Avraham O, Tsaban T, Ben-Aharon Z, et al. Protein language models can capture protein quaternary state. BMC Bioinf, 2023, 24: 433
105 Hou X, Wang Y, Bu D, et al. EMNGly: predicting N-linked glycosylation sites using the language models for feature extraction. Bioinformatics, 2023, 39: btad650
106 Xu M, Yuan X, Miret S, et al. ProtST: multi-modality learning of protein sequences and biomedical texts. In: Proceedings of the 40th International Conference on Machine Learning, 2023. 38749–38767
107 Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res, 2015, 43: D1049–D1056
108 Brandes N, Ofer D, Peleg Y, et al. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioin-

formatics, 2022, 38: 2102–2110
109 Meier J, Rao R, Verkuil R, et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021. 34: 29287–29303
110 Marquet C, Heinzinger M, Olenyi T, et al. Embeddings from protein language models predict conservation and variant effects. Hum Genet, 2022, 141: 1629–1647
111 Strodthoff N, Wagner P, Wenzel M, et al. UDSMProt: universal deep sequence models for protein classification. Bioinformatics, 2020, 36: 2401–2409
112 Madani A, McCann B, Naik N, et al. ProGen: language modeling for protein generation. 2020. ArXiv:2004.03497
113 Nijkamp E, Ruffolo J A, Weinstein E N, et al. ProGen2: exploring the boundaries of protein language models. Cell Syst, 2023, 14: 968–978
114 van Kempen M, Kim S S, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol, 2024, 42: 243–246
115 Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data. Nature, 2024, 631: 755–759
116 Fenoy E, Edera A A, Stegmayer G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. Brief Bioinf, 2022, 23: bbac232
117 Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature, 2021, 596: 590–596
118 Fang Y, Xu F, Wei L, et al. AFP-MFL: accurate identification of antifungal peptides using multi-view feature learning. Brief Bioinf, 2023, 24: bbac606
119 Wang M, Patsenker J, Li H, et al. Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. Nucleic Acids Res, 2024, 52: 548–557
120 Ruffolo J A, Gray J J, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. 2021. ArXiv:2112.07782
121 Xu S, Shen L, Zhang M, et al. Surface-based multimodal protein-ligand binding affinity prediction. Bioinformatics, 2024, 40: btae413
122 Wang X, Yin X, Jiang D, et al. Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. Nat Commun, 2024, 15: 7348
123 Geffen Y, Ofran Y, Unger R. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. Bioinformatics, 2022, 38: ii95–ii98
124 Mirabello C, Wallner B, Zhang Y. rawMSA: end-to-end deep learning using raw multiple sequence alignments. Plos One, 2019, 14: e0220182
125 Zhou H, Yang Y, Shen H B, et al. Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. Bioinformatics, 2017, 33: 843–853
126 Alamdari S, Thakkar N, van den Berg R, et al. Protein generation with evolutionary diffusion: sequence is all you need. bioRxiv, 2023. doi: 10.1101/2023.09.11.556673
127 Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage? In: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2024. 36: 12345–12358
128 Misteli T, Spector D L. Applications of the green fluorescent protein in cell biology and biotechnology. Nat Biotechnol, 1997, 15: 961–964
129 Kulmanov M, Guzmán-Vega F J, Roggli P D, et al. Protein function prediction as approximate semantic entailment. Nat Mach Intell, 2024, 6: 220–228
130 Shanker V R, Bruun T U J, Hie B L, et al. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. Science, 2024, 385: 46–53
131 Chen B, Cheng X, Li P, et al. xTrimoPGLM: unified 100B-scale pre-trained transformer for deciphering the language of protein. 2024. ArXiv:2401.06199
132 Kortemme T. De novo protein design From new structures to programmable functions. Cell, 2024, 187: 526–544
133 Yeung W, Zhou Z, Li S, et al. Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings. Brief Bioinf, 2023, 24: bbac599
134 Nallapareddy V, Bordin N, Sillitoe I, et al. CATHe: detection of remote homologues for CATH superfamilies using embeddings from protein language models. Bioinformatics, 2023, 39: btad029
135 Suzek B E, Huang H, McGarvey P, et al. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007, 23: 1282–1288
136 Suzek B E, Wang Y, Huang H, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 2015, 31: 926–932
137 Finn R D, Mistry J, Schuster-Böckler B, et al. Pfam: clans, web tools and services. Nucleic Acids Res, 2006, 34: D247–D251
138 Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res, 2000, 28: 45–48
139 Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res, 2022, 50: D439–D444
140 Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods, 2019, 16: 603–606
141 Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun, 2018, 9: 2542
142 Bateman A, Martin M J, Orchard S, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res, 2023, 51: D523–D531
143 Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25–29
144 McDonald A G, Boyce S, Tipton K F. ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res, 2009, 37: D593–D597
145 Wu Y, Yi X, Tan Y, et al. A PLMs based protein retrieval framework. 2024. ArXiv:2407.11548
146 Thul P J, Lindskog C. The human protein atlas: a spatial map of the human proteome. Protein Sci, 2018, 27: 233–244
147 Szklarczyk D, Gable A L, Nastou K C, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res, 2021, 49: D605–D612
148 Jin M, Xue H, Wang Z, et al. ProLLM: protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. bioRxiv, 2024. doi: 10.1101/2024.04.18.590025
149 Orengo C A, Michie A D, Jones S, et al. CATH—a hierarchic classification of protein domain structures. Structure, 1997, 5: 1093–1109
150 Rao R, Bhattacharya N, Thomas N, et al. Evaluating protein transfer learning with TAPE. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019. 32: 9689–9701
151 Alkuhlani A, Gad W, Roushdy M, et al. PTG-PLM: predicting post-translational glycosylation and glycation sites using protein language models and deep learning. Axioms, 2022, 11: 469
152 Notin P, Kollasch A W, Ritter D, et al. ProteinGym: large-scale benchmarks for protein design and fitness prediction.

bioRxiv, 2023. doi: 10.1101/2023.12.07.570727
153 Dallago C, Mou J, Johnston K E, et al. FLIP: benchmark tasks in fitness landscape inference for proteins. In: Proceedings of Neural Information Processing Systems, 2021
154 Xu H, Zhou J, Lin S, et al. PLMD: an updated data resource of protein lysine modifications. J Genet Genomics, 2017, 44: 243–250
155 Zhang W, Tan X, Lin S, et al. CPLM 4.0: an updated database with rich annotations for protein lysine modifications. Nucleic Acids Res, 2022, 50: D451–D459
156 Mardikoraem M, Wang Z, Pascual N, et al. Generative models for protein sequence modeling: recent advances and future directions. Brief Bioinf, 2023, 24: bbad358
157 Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. In: Proceedings of Neural Information Processing Systems, 2022. 35: 22199–22213
158 Chae Y, Davidson T. Large language models for text classification: from zero-shot learning to instruction-tuning. Sociol Methods Res, 2025. doi: 10.1177/00491241251325243
159 Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM Comput Surv, 2021, 54: 1–35
160 Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data, 2016, 3: 160035
161 Pan X, Zhang M, Ji S, et al. Privacy risks of general-purpose language models. In: Proceedings of IEEE Symposium on Security and Privacy (SP), San Francisco, 2020. 1314–1331
162 Wagner A. AI predicts the effectiveness and evolution of gene promoter sequences. Nature, 2022, 603: 399–400
163 Yu Y, Buchanan S, Pai D, et al. White-box transformers via sparse rate reduction. In: Proceedings of Neural Information Processing Systems, 2023. 36: 9422–9457
164 Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell, 2020, 2: 56–67
165 Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016. 1135–1144
166 Jin Z, Liu J, Lyu Z, et al. Can large language models infer causation from correlation? 2023. ArXiv:2306.05836
167 Scholkopf B. Causality for machine learning. In: Probabilistic and Causal Inference: The Works of Judea Pearl. New York: Association for Computing Machinery, 2022. 765–804
168 Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. 2023. ArXiv:2312.11805
169 Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat Methods, 2024, 21: 1462–1465
170 Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinf, 2022, 23: bbac409
171 Ille A M, Markosian C, Burley S K, et al. Generative artificial intelligence performs rudimentary structural biology modeling. Sci Rep, 2024, 14: 19372
172 Wang L, Chen X, Deng X W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. npj Digit Med, 2024, 7: 41
173 Zhang Y J, Luo Z, Sun Y, et al. From beasts to bytes: revolutionizing zoological research with artificial intelligence. Zool Res, 2023, 44: 1115–1131
174 Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature, 2024, 630: 181–188
175 Wang C W, Chang C C, Khalil M A, et al. Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer. Sci Data, 2022, 9: 25
176 Chen R J, Lu M Y, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans Med Imag, 2020, 41: 757–770
177 Chen R J, Lu M Y, Williamson D F K, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell, 2022, 40: 865–878
178 Wang T, Geng J, Zeng X, et al. Exploring causal effects of sarcopenia on risk and progression of Parkinson disease by Mendelian randomization. npj Parkinsons Dis, 2024, 10: 164
179 Wang T, Yan Z, Zhang Y, et al. postGWAS: a web server for deciphering the causality post the genome-wide association studies. Comput Biol Med, 2024, 171: 108108
180 Rothman D. Transformers for Natural Language Processing: Build, Train, and Fine-Tune Deep Neural Network Architectures for NLP With Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4. Birmingham: Packt Publishing Ltd., 2022. 1–500
181 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929
182 Anonymous Authors. Feature re-embedding: towards foundation model-level performance in computational pathology. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, 2024. 11343–11352
183 Huang Z, Bianchi F, Yuksekgonul M, et al. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med, 2023, 29: 2307–2316