Large Language Models for Molecular Biology: Bridging Computational Advances and Biomolecular Insights

Anonymous ACL submission

Abstract

Large language models (LLMs) are transforming numerous sectors and are increasingly being explored to advance molecular biology by enabling computational analysis of biological language. However, the grammatical and semantic complexities of biomolecules present challenges for LLMs. This survey explores three key strategies to bridge this gap: (1) biological LLMs, pretrained on biological language to capture unimodal representation or multimodal (i.e., sequence-structure) relationships, (2) post-training adaptations, which refine natural LLMs through instruction-tuning or retrieval-augmented generation, and (3) multimodal LLMs, which is capable of jointly processing biological and natural languages. In this work, we highlight the potential of multimodal LLMs that integrate biomolecular data, general and scientific literature knowledge to enhance biological language processing, thus accelerating molecular biology research while addressing the aforementioned challenges.

1 Introduction

Biomolecules, including proteins, nucleic acids and small molecules, are fundamental to cellular functions and homeostasis. Understanding and reasoning over their sequences, structures, and functions is key to deciphering biological processes. Additionally, biomolecular generation enables the design of molecules with tailored properties. Their efficacy depends not only on intrinsic properties but also on interactions within biological systems. Therefore, elucidating biomolecular interactions enhances our understanding of disease mechanisms (Sebastian-Leon et al., 2014) and drive innovations in novel therapeutic target identification (Nowell et al., 2023), treatment optimation (Negishi et al., 2024), personalized medicine (Goetz and Schork, 2018), biomarker discovery (Ou et al., 2021), and biomolecular engineering (Victorino da Silva Amatto et al., 2022). However, computational analysis of biomolecules remains challenging due to their high-dimensional nature, intricate interactions, and diverse functions. Recent achievements in artificial intelligence, particularly large language models (LLMs), offer promising solutions by leveraging large-scale biomolecular data to extract meaningful representations and relationships.

Advances in natural LLMs have inspired researchers to apply language modeling techniques to biological sequences, treating nucleotide, amino acid sequences, and molecular representations as structured data. This perspective has led to the development of biological LLMs, including Evo (Nguyen et al., 2024b), ESM-2/ESM-fold (Lin et al., 2022), ESM3 (Hayes et al., 2025) and NatureLM (Xia et al., 2025), which excel in protein structure prediction, molecular property prediction, and DNA sequence design, etc. However, the biological LLMs, primarily pretrained on domain-specific corpora, have limited coverage of general and scientific literature knowledge, restricting their performance on complex biological language tasks that require cross-disciplinary integration. To address this, researchers have explored adapting natural LLMs through instruction tuning (Xiao et al., 2024a) or retrieval-augmented generation (RAG) (Lin et al., 2024), showing some success in biological language processing. Still, natural LLMs, primarily pretrained on plain text, struggle with more complex biomolecular data modalities such as 2D molecular graphs and 3D protein structures.

Given that natural LLMs cannot fully capture the complexity of biomolecules, researchers have developed multimodal LLMs to integrate diverse biological data. These models process sequences, structures, and texts (functions) simultaneously while incorporating domain knowledge to improve biological understanding, reasoning and generation. Recent advances, such as Evola (Zhou

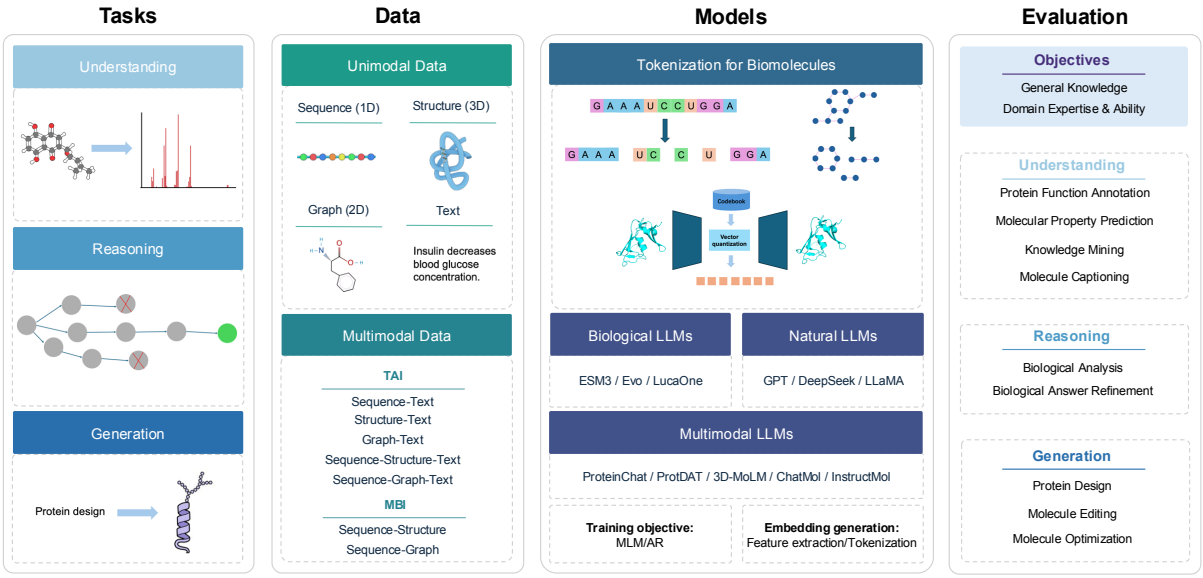

Figure 1: Overview of LLMs for molecular biology, encompassing models, data, tasks, and evaluation. TAI: Text Aligned Integration. MBI: Multiple Biomolecules Integration.

et al., 2025a), highlight the potential of modeling biomolecules in 3D space while leveraging textual data for biological inference. This trend underscores the growing role of multimodal LLMs in computational biology, paving the way for next-generation biological computation models.

As shown in Figure 1, we categorize existing works by biomolecular data types, analyzing their modalities and key applications. We also highlight emerging methodology to construct LLMs for molecular biology. This survey offers a comprehensive research dimensions and potential applications of LLMs to transform molecular biology research.

2 Natural and biological language

Natural language, such as English, French, Chinese, and Japanese, are a communication system developed by humans to express thoughts, ideas, and information. It evolves naturally and are characterized by defined syntax, semantics, and pragmatics. Similarly, biological language refers to structured, information-rich encoding systems that regulate biological processes, encompassing protein sequences and structures, RNA and DNA sequences, molecular strings such as SMILES, and cellular signaling pathways. As shown in Table 5 of Appendix A, despite their distinct origins and purposes, both natural and biological languages encode complex, structured information, enabling communication within their respective domains. However, biological language differs fundamentally from natural language in grammars and semantics. While natural language derives meaning from linguistic and contextual relationships, biological language primarily encodes functions through structural properties and biomolecular interactions.

Although the development of NLP techniques has also brought promising applications in biological language processing, significant challenges remain. Natural LLMs lack a fundamental understanding of biomolecular structures and interactions. Unlike natural language, where semantic meaning arises from word relationships and syntax, biological language is governed by biophysical and biochemical principles that cannot be fully captured by sequence-based models alone. **Overall, the uniqueness and challenges of biological language are as follows:**

- **Lack of explicit grammar and semantics**

  Natural language follows well-defined grammatical rules, while biological language lack fixed syntax rules. Biomolecular functions often require experimental validation rather than relying solely on sequence analysis. For example, a DNA fragment may encode a protein or function as a regulatory region, but its true role is difficult to determine based solely on sequence analysis.

- **Highly structured and multimodal nature**

  Unlike natural language, biological language is highly structured and multimodal, encom-

passing sequences, structures, and biophysical and biochemical properties. For example, cellular signaling involves spatial, temporal, and multi-layered regulation. As a result, sequence-based models often fail to capture these complexities, requiring integration with additional components such as graph neural networks and multimodal connectors. While natural LLMs excel in pattern recognition, they struggle with structural modeling and causal inference in biological contexts. They fail to capture dependencies between protein structures and functions or between genetic mutations and phenotypic effects, as biological systems are highly nonlinear, dynamic, and context-dependent. Addressing these challenges requires multimodal learning and causal modeling to enhance LLMs for molecular biology.

Therefore, although natural LLMs can be partially applied to biological language tasks, truly leveraging LLMs to solve biological language tasks requires the integration of general knowledge and specialized domain expertise and abilities by integrating diverse biological data modalities.

3 Biological language tasks

Recent advancements in bioinformatics and computational biology have increasingly focused on applying LLMs to biological language tasks that can be broadly categorized into three main categories: **understanding, reasoning, and generation**. These three categories cover a wide range of challenges in computational biology, and applying LLMs to these tasks has the potential to transform molecular biology research, enhancing our understanding of biomolecular systems and accelerating scientific discoveries.

Understanding. This category encompasses tasks related to parsing and interpreting biological sequences, structures, and their functional implications. Its primary objective is to extract meaningful insights from biomolecules and uncover the fundamental principles that govern biological processes. Applications includes protein functions or active sites prediction from structures, gene functions or non-coding regions identification from DNA sequences, cell type prediction from single-cell sequencing data and disease diagnosis from multi-omics data.

Reasoning. This category involves tasks that require higher-order thinking, such as predicting complex biological relationships and causal mechanisms. It aims to generate new knowledge and infer previously unknown biological processes by leveraging existing data. These tasks can be broadly classified into predictive modeling and causal inference. Applications include novel therapeutic target identification from gene-protein-disease networks, gene-phenotype association prediction from genetic variations and phenotypic traits, identifying critical genes or molecular interactions from biological pathways, and disease mechanism inference from multi-omics data.

Generation. This category focuses on generating biologically meaningful molecules, including sequences and structures. It targets to design novel molecules with tailored properties or functions. Applications include protein design with specific functions, antibody design with high affinity and specificity, novel chemical structures generation for drug discovery, gene editing tool generation creating precise and efficient CRISPR sequences, and antigen epitopes design based on pathogen data.

4 Biological data modalities

In computational biology and artificial intelligence, data fall into two categories: unimodal data and multimodal data (see Table 1). Unimodal data consists of a single input type, while multimodal data integrate multiple types, reflecting the complexity of real-world biological systems. Understanding the distinction between unimodal and multimodal data is essential for developing advanced models that address the complexities of molecular biology research.

Table 1: Comparison of unimodal and multimodal data in molecular biology. TAI: Text Aligned Integration. MBI: Multiple Biomolecules Integration.

Modality	Fusion method	Examples
Unimodal	-	Sequence, Structure, Graph, Text
Multimodal	TAI	Sequence-Text, Structure-Text, Graph-Text, Sequence-Structure-Text, Sequence-Graph-Text
	MBI	Sequence-Structure, Sequence-Graph

Unimodal data serves as the foundation for many applications in molecular biology. Examples include protein sequences, where models predict function or structure based solely on amino acid sequences, and biological graphs, which represent molecular interactions or networks. Similarly, biological structures, such as 3D protein conformations, are used to understand biophysical properties

or ligand binding. Beyond biological data, general text and biological text, such as research articles or clinical reports, provide valuable information for tasks like named entity recognition or knowledge mining.

Multimodal data aims to enable comprehensive understanding of complex biomolecules by integrating multiple biological data modalities, such as sequence, structure, and text (function). Cross-modal fusion presents new opportunities to bridge biomolecular insights with broader general and scientific literature knowledge. As multimodal approaches capture the multifaceted nature of biomolecular data, they are increasingly utilized in molecular biology research. The primary forms of multimodal data fall into two categorizes: Text Aligned Integration (TAI) and Multiple Biomolecule Integration (MBI).

TAI uses text to enhance the understanding of biological modalities, such as learning protein functions or gene regulation from scientific literature. **Sequence-Text** combines sequences with text, linking molecular information to functional descriptions for better understanding of genetic sequences and biological functions. **Structure-Text** merges structures with text to connect spatial features with biological interpretations, helping to relate 3D structures to protein functions. **Graph-Text** integrates graph data, like molecular topologies, with text, supporting drug discovery by associating molecular graphs with literature insights. **Sequence-Structure-Text** combines sequences, 3D structures, and text, providing a holistic view of how sequence variations affect biological functions. **Sequence-Graph-Text** connects sequence data with graph representations and text, aiding in drug development and gene-disease associations by contextualizing sequence variations.

MBI focuses on the integration of biomolecular data itself, utilizing the multimodal characteristics of sequences, structures, and graphs to improve predictive capabilities. **Sequence-Structure** integration links biomolecular sequences with their corresponding 3D conformations. This allows models to capture how structural properties emerge from linear sequences, facilitating a better understanding of how sequence variations influence protein folding and function. **Sequence-Graph** incorporates sequence information with graph-based representations, enabling a more comprehensive understanding of biological mechanisms by embedding sequence-derived features within graph features. This can enhance applications such as protein-protein interaction prediction and molecular pathway analysis.

5 Model methodology

5.1 Training objective

Masked language modeling (MLM) is a commonly used self-supervised pretraining approach, in which certain input tokens are masked, and the model is trained to predict them using contextual information. Through this method, the model can learn bidirectional representations, improving its performance on tasks requiring deep contextual understanding. In ESM3, MLM integrates multiple modalities, such as sequence, structure, and text, with the filling of each masked token conditioned on various modalities. This interaction process enables ESM3 to capture the residue-level evolutionary path dependencies within the protein space. This includes modeling the feasible sequence space under specific protein conformational constraints and structural selection driven by functional constraints, providing a deeper understanding of protein evolution. From the perspective of protein design, the mechanism of ESM3 allows it to accept flexible combinations of modalities as input for designing novel proteins.

Autoregressive (AR) models, like GPT, generate tokens sequentially, conditioning each on the previous ones. This unidirectional approach is ideal for tasks requiring sequential coherence, such as text completion. This autoregressive method also allows EVO to capture sequential dependencies in genomic data, ensuring more accurate and coherent generation of long-range genetic structures. In genomes, functional elements like promoters and enhancers are distributed with a sequential relationship, making AR particularly well-suited for modeling the sequential semantic information of genomic sequences.

5.2 Embedding generation

Data embedding generation involve two key methods: feature extraction and tokenization. Both methods are essential for converting raw data into representations that can be effectively utilized by LLMs for diverse downstream tasks.

Feature extraction aim to process raw data to extract meaningful representations that capture essential information. Using pre-trained encoders, the raw data such as biomolecular sequences,

graphs, and structures are transformed into high-dimensional feature vectors. These features serve as abstract embeddings, which can then be used for further analysis, prediction, or understanding of biomolecular properties. This method enables a deeper understanding of the raw data by converting raw information into a more manageable and interpretable format.

Tokenization treats biomolecular sequences as discrete symbols, where individual residues, nucleotides, functional motifs, or atoms are tokenized. This strategy leverages techniques such as k-mer modeling, subword tokenization, to segment sequences into meaningful units, allowing for efficient learning of sequence patterns. In addition, biomolecular structures can be tokenized via vector quantization (VQ). Unlike sequence data, structural representations are inherently continuous, making direct tokenization challenging. VQ addresses this by discretizing 3D coordinates into a predefined codebook, capturing essential geometric and topological features. This enables structural data to be processed similarly to sequence data while preserving critical spatial information.

5.3 Model architecture

In this survey, LLMs fall into two categories based on language types. **Biological LLMs (5.3.1)** are trained on biomolecular sequences, graphs and structures or their integration, capturing domain-specific patterns for biological language tasks. While effective in processing biological language, they lack general knowledge. **Natural LLMs (5.3.2)** excel in understanding and generating human language but struggle with biomolecular complexity. Adaptations such as instruction tuning help them handle biological tasks but remain limited in tasks related to biomolecular graphs and structures. **Multimodal LLMs (5.3.3)** bridge biological and natural language by integrating sequences, graphs structures, and textual data, enabling cross-modal understanding, reasoning and generation. Table 2 shows recent progress of LLMs in molecular biology, encompassing single cell, protein, DNA, RNA, small molecules, etc.

5.3.1 Biological large language models

Biological LLMs are computational models specifically trained on large-scale biological data, including amino acid sequences, nucleotide sequences, SMILES representations, and single-cell sequencing data. Examples include ESM3 (Hayes et al., 2024), Nucleotide Transformer (Dalla-Torre et al., 2024), and Evo (Nguyen et al., 2024a). These models build upon advancements in language modeling techniques to learn the unique properties of biological language (Li et al., 2021), aiming to uncover meaningful insights encoded within them. Depending on the type of biomolecular data they process, biological LLMs can be further categorized into the following five model classes (see Appendix B), each tailored to address specific challenges in their respective domains.

5.3.2 Natural large language models

Natural and biological LLMs, although operating in distinct domains, share fundamental principles for processing sequential data. Over the past decade, language models have transformed NLP (Mikolov et al., 2013; Pennington et al., 2014), with key breakthroughs driven by transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018; Radford et al., 2019). Their attention mechanisms enable efficient modeling of long-range dependencies, leading to increasingly powerful models like GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022), and GPT-4 (Achiam et al., 2023), DeepSeek (Liu et al., 2024a), excelling in text generation, translation, and question-answering. Natural LLMs have been adapted to biological language tasks via instruction fine-tuning (Fang et al., 2024a), but they struggle with more complex biomolecular data, such as 2D molecular graphs and 3D protein structures. This limitation has spurred the development of multimodal LLMs (MLLMs), which integrate text with diverse biomolecular data types to bridge natural and biological languages.

5.3.3 Multimodal large language models

With the rapid advancement of natural LLMs, MLLMs, which process and integrate multimodal data, have gained significant attention, such as BLIP-2 (Li et al., 2023), Kosmos-1 (Huang et al., 2023), and Llava (Liu et al., 2024c). These models not only enhance performance in language-related tasks, but also advance the capability of models to understand and generate information that spans both abstract and physical domains. MLLMs have shown promise in a wide range of tasks, such as text-to-image/video generation, video analysis, and audio-visual understanding, which are essential steps towards artificial general intelligence (AGI).

These developments are pushing the boundaries of machine understanding by facilitating more nuanced interactions with the real world. In this survey, we term MLLMs as those trained on integration of textual data with heterogeneous biomolecular data (i.e., using TAI for data integration), enabling deeper insights into complex biomolecular interactions. In molecular biology, MLLMs leverage mainstream multimodal architectures for modality alignment, such as projection-based (e.g., Llava) and Q-Former-based (e.g., BLIP-2) models. These architectures align different modalities by projecting them into a shared space or using query encoders and cross attention mechanisms to extract and fuse key biological information.

6 Tokenization for biomolecules

Tokenization plays a crucial role in biological language processing, akin to its importance in natural language processing, as it directly affects a model's ability to interpret input and overall performance (Pei et al., 2024). Tokenization techniques for biomolecular data have evolved to address diverse modalities. For 1D biomolecular sequences, methods like K-mer tokenization segment sequences into fixed-length substrings, while subword methods focus on biologically relevant motifs. For 2D graphs and 3D structures, tokenization incorporates biomolecular topologies and spatial relationships. Approaches such as graph-based tokenizers and vector quantization (VQ) capture graph and structure features, enabling models to effectively represent 2D and 3D biomolecular data.

6.1 1D sequence tokenization

K-mer tokenization is a widely used technique for processing biological sequences, particularly in genomics and proteomics. This method involves segmenting sequences into fixed-length overlapping or non-overlapping substrings, known as K-mer, where K represents the length of each substring. For example, in DNA sequences, a 3-mer tokenization would break the sequence "ATGCGT" into ["ATG", "TGC", "GCG", "CGT"].

Byte-pair-encoding (BPE) is a renowned subword tokenization method that constructs a variable-length vocabulary by repeatedly merging the most frequent adjacent symbol pairs. It has been widely adopted in NLP to efficiently handle rare and out-of-vocabulary words. Its effectiveness in managing long sequences and rare token occurrences, has led to applications in biomolecular data. For example, DNABERT-2 (Zhou et al., 2023) merges frequent nucleotide pairs and genome segments to enhance genomic sequence representation by capturing local and long-range dependencies. Despite improving tokenization efficiency, BPE has limitations in biological contexts. Unlike natural language, where subword units carry semantic meaning, biological sequences lack explicit word boundaries, complicating meaningful tokenization. BPE's frequency-based merging may overlook functional motifs or structural elements, potentially leading to biologically irrelevant segmentations. Additionally, frequent patterns in one dataset may not generalize across species or sequence contexts. Addressing these challenges requires biologically informed tokenization that incorporate structural, functional, or evolutionary constraints.

Hierarchical encoding is a technique that represents biomolecules at multiple levels of abstraction to preserve both local and global contextual information. Unlike conventional tokenization methods that operate at a single level (e.g., residue-level and atom-level), hierarchical encoding introduces multiple layers of representation to capture the structural and functional complexity of biological data. For example, HELM (Yazdani-Jahromi et al., 2024) encodes mRNA at multiple levels, such as 6-mer and codon-level tokenization, to seize the biological significance of mRNA sequences.

Specialized vocabulary refers to a tokenization approach where domain-specific token dictionaries are designed for biomolecular data. Instead of relying on purely statistical subword segmentation methods, specialized vocabulary approaches incorporate domain knowledge to define biologically meaningful tokens, such as amino acid motifs, codons, or functional domains, to enhance model interpretability and performance in biomolecular tasks. The work (Ai and Kavuluru, 2023) customized tokenization way and specialized dictionaries for biomolecular sequences, providing empirical evidence of their effectiveness.

6.2 2D graph and 3D structure tokenization

Graph-based tokenizers represent biomolecules as graphs, where atoms are nodes and bonds are edges, capturing spatial relationships and connectivity. Tokenization involves encoding node features (e.g., atom type, charge) and edge features (e.g., bond type, distance) into embeddings for pro-

Table 2: Recent research progress of large language models for molecular biology

Model/Method	Biomolecule type	Data modality	Key takeaway
ESMFold (Lin et al., 2023b)	Protein	Sequence-Structure	Protein structure prediction
ESM3 (Hayes et al., 2025)	Protein	Sequence-Structure-Text	Multimodal, Large-scale token level pretraining
AIDO.RAGFold (Li et al., 2024c)	Protein	Sequence-Structure	Protein structure prediction, RAG
EVO (Nguyen et al., 2024b)	DNA	Sequence	Genome generation
Nucleotide Transformer (Dalla-Torre et al., 2024)	DNA	Sequence	Genome foundation model
GenomeOcean (Zhou et al., 2025b)	DNA	Sequence	Large-Scale, Metagenomic
MegaDNA (Shao and Yan, 2024)	DNA	Sequence	Long-context generative model
scGPT (Cui et al., 2024)	RNA	Sequence	Single-cell foundation model, Multi-omic
LucaOne (He et al., 2024b)	DNA, RNA, Protein	Sequence	Large-scale genome foundational model
NatureLM (Xia et al., 2025)	RNA, Protein, Small molecules	Sequence	Sequence-based molecule foundation model
InstructProtein (Wang et al., 2023)	Protein	Sequence-Text	Bidirectional generation of protein sequence and language, Knowledge graph-based instruction
HelixProtX (Chen et al., 2024)	Protein	Sequence-Structure-Text	Transformation between protein sequences, structures, and textual descriptions
SEPIT (Wu et al., 2024)	Protein	Sequence-Structure-Text	Protein function prediction, Mixture of experts (MoE), Instruction tuning
ProtT3 (Liu et al., 2024e)	Protein	Sequence-Text	Protein Understanding, Q-Former Cross-modal contrastive learning
ProLLM (Jin et al., 2024)	Protein	Sequence-Text	Protein-protein interaction prediction Protein chain of thought
PROTLLM (Zhuo et al., 2024)	Protein	Sequence-Text	Interleaved protein-text dataset Dynamic protein mounting
ProLLaMA (Lv et al., 2024)	Protein	Sequence-Text	Protein sequence generation and understanding Protein vocabulary pruning
P-LLMs (Zeinalipour et al., 2024)	Protein	Sequence-Text	Tokenizer retraining, Adaptation to small datasets
ProtDAT (Guo et al., 2024)	Protein	Sequence-Text	De novo protein design Multimodal cross-attention
BioM3 (Praljak et al., 2024)	Protein	Sequence-Text	Protein domain design, Contrastive learning
ProteinGPT (Xiao et al., 2024b)	Protein	Sequence-Structure-Text	Protein property prediction Linear projection
ProteinChat (Huo et al., 2024)	Protein	Sequence-Text	Interactive refinement, Adaptor, LoRA
ProtChatGPT (Wang et al., 2024a)	Protein	Sequence-Structure-Text	Interactive conversations about protein structures Multi-level protein-language alignment
Protein Captioning (Zhang et al., 2024b)	Protein	Sequence-Text	Protein captioning, Conversational interaction
Evola (Zhou et al., 2025a)	Protein	Sequence-Structure-Text	Direct preference optimization, RAG AI-generated data
TourSynbio (Shen et al., 2024)	Protein	Sequence-Text	Protein engineering, Mutation analysis, Agent
PQA (Carrami and Sharifzadeh, 2024)	Protein	Sequence-Text	Protein question answering, Soft prompts Gated cross-attention
RSA (Ma et al., 2024)	Protein	Sequence-Text	RAG, LLM agents
MolT5 (Edwards et al., 2022)	Small molecules	Sequence-Text	Molecule captioning and text-based molecule generation
InstructMol (Cao et al., 2023)	Small molecules	Sequence-Graph-Text	Assistants in molecular research Multimodal instruction-tuning
MolCA (Liu et al., 2023)	Small molecules	Sequence-Graph-Text	Gragh-text alignment Q-Former, LoRA
STRUCTCOT (Jang et al., 2024)	Small molecules	Sequence-Text	Structure-aware, Chain-of-thought
GIT-Mol (Liu et al., 2024d)	Small molecules	Graph-Text	Multimodal LLMs that integrates graph, text, and image
ICMA (Li et al., 2024b)	Small molecules	Sequence-Graph-Text	In-context molecule tuning Hybrid context retrieval
3D-MoLM (Li et al., 2024d)	Small molecules	Sequence-Structure-Text	Connects LLMs and 3D molecular encoder Q-Former
ChemLLM (Zhang et al., 2024a)	Small molecules	Sequence-Text	Chemistry, Instruction tuning Benchmark for chemistry tasks
ChemDFM (Zhao et al., 2024b)	Small molecules	Sequence-Text	Chemical literature, Instruction tuning
nach0 (Livne et al., 2024)	Small molecules	Sequence-Text	Pre-trained on scientific literature and molecular data, Drug discovery
BioT5 (Pei et al., 2023)	Small molecules, Protein	Sequence-Sequence-Text	integrates molecular string SELFIES and contextual knowledge from literature
BioMedGPT (Luo et al., 2023)	Small molecules, Protein	Sequence-Text	Incremental training on literature, Biomedical QA
InstructBioMol (Zhuang et al., 2024)	Small molecules, Protein	Sequence-Structure-Text	Understanding and design of biomolecules Motif-Guided multimodal module
ChatNT (Richard et al., 2024)	Protein, DNA, RNA	Sequence-Text	Multimodal LLMs for DNA, RNA, and protein
LangCell (Zhao et al., 2024a)	RNA	Sequence-Text	Understanding of cell identity Language-cell pre-training
CHATCELL (Fang et al., 2024b)	Gene	Text	Single-cell analysis
GeneRAG (Lin et al., 2024)	Gene	Text	RAG, Gene, LLMs
BioRAG (Wang et al., 2024b)	Gene, Protein	Text	RAG

cessing by graph neural networks (GNNs). This approach preserves the biomolecular topology, making it effective for tasks like molecular property prediction, drug discovery, and protein-ligand interactions. However, challenges include managing large graphs, ensuring efficient graph convolution

operations, and maintaining interpretability.

Vector quantization (VQ) methods construct vocabularies for continuous biomolecular data, especially 3D protein structures (see Table 3). VQ discretizes continuous representations into a codebook, serving as tokens for downstream tasks. This results in compact and meaningful vocabularies that retain essential biomolecular features while reducing complexity. However, VQ methods face challenges in biomolecular applications. Discretization may lose critical structural and functional details, as biomolecules exhibit hierarchical and multi-scale properties that may not be well-captured by a fixed set of quantized codes. Additionally, determining an optimal codebook size is crucial—too few codes may oversimplify complex biomolecular representations, while too many introduce sparsity and inefficiency. Furthermore, VQ methods struggle to generalize across diverse biological contexts due to species- and condition-specific variability. Addressing these limitations may require biologically informed constraints, adaptive quantization, or hybrid approaches balancing continuous and discrete representations.

Table 3: Vector quantization (VQ) methods.

Method	Key takeaway
VQPL (Gao et al., 2023)	Quantized protein language
ProTokens (Lin et al., 2023a)	Probabilistic tokenization
bio2token (Liu et al., 2024b)	Large molecular structures, Mamba, State space model architecture
LPS (Gaujac et al., 2024)	Vector-quantized autoencoder, Multi-modal integration
FoldToken (Gao et al., 2024d)	Soft conditional vector quantization
FoldToken2 (Gao et al., 2024b)	Equivariant structures, Vector-quantized compressor
FoldToken3 (Gao et al., 2024a)	Light-weight, efficient tokenization
FoldToken4 (Gao et al., 2024c)	Hierarchical multiscale, Token mixing

7 Evaluation of LLMs in molecular biology

The evaluation of LLMs in molecular biology requires specialized benchmarks that assess their ability to understand, generate, and reason over biological data. Table 4 offers an overview of key datasets and benchmarks designed for evaluating LLMs in life sciences, covering various task such as molecular understanding, generation, and reasoning. These benchmarks focus on different aspects of molecular biology. By leveraging these diverse benchmarks, researchers can systematically analyze the advantages and shortcomings of LLMs in biological language tasks, guiding future improvements and adaptations for more effective applications in molecular biology research.

Table 4: Datasets and benchmarks for evaluation of LLMs in life sciences.

Model/Method	Task scope	Key takeaway
Bioinfo-Bench (Chen and Deng, 2023)	Understanding	LLMs for bioinformatics
SciEval (Sun et al., 2023)	Understanding	Scientific research, Multi-level
LLMaMol (Zhong et al., 2024)	Understanding	Geometric structure of molecules
MoleculeQA (Lu et al., 2024)	Understanding	Factual evaluation
MolCap-Arena (Edwards et al., 2024)	Understanding	Molecular property prediction Molecule caption
OPI (Xiao et al., 2024a)	Understanding	Annotation prediction, Sequence understanding, Knowledge mining
Biology Instructions (He et al., 2024a)	Understanding, Reasoning	Multi-omics, Large-scale
Mol-Instructions (Fang et al., 2024a)	Understanding, Generation	Protein design, Instruction tuning
TOMG-Bench (Li et al., 2024a)	Generation	Molecule editing, Molecule optimization
ProteinBench (Ye et al., 2024)	Generation	Multi-metrics, Antibody design Inverse folding

8 Conclusion and future work

LLMs are transforming molecular biology by enabling biological language computation, yet challenges persist due to the grammar and semantic complexities of biomolecules. This survey examines three key strategies to address these challenges: (1) biological LLMs trained on biomolecular data, (2) post-training adaptations like instruction-tuning, and (3) multimodal LLMs integrating sequences, structures, and functions. Among these, multimodal LLMs show the greatest promise, as they unify biomolecular data, general knowledge, and scientific literature to enhance biological language processing and accelerate scientific discoveries. Yet, challenges like biologically informed tokenization and cross-domain generalization, still remain. Future research directions suggested in this work include: (1) developing adaptive tokenizers (Yan et al., 2024b; Shen et al., 2025) for biomolecules, which dynamically adjust token granularity based on input data characteristics, optimizing tokenization through variable token lengths and context-specific rules, making them particularly suited for domains like biological language modeling with highly variable structures; (2) optimizing the integration of specialized biomolecular encoders and LLMs to align biological and natural languages; (3) developing comprehensive benchmarks for multimodal LLMs in biomolecular tasks, improving model generalization across various downstream tasks of interest to domain experts.

Limitations

This survey focuses on LLMs in molecular biology, specifically targeting biomolecular sequences, structures, graphs and functions (texts). They are key aspects of molecular biology that have gained attention in recent years. The LLMs listed in this survey do not address other data types, such as cell images or experimental data. With the development of molecular biology research, future surveys could cover more advanced LLMs that incorporate a broader range of data types and explore their applications in molecular biology.

Ethics Statement

To the best of our knowledge, there are no ethical concerns associated with the LLMs listed in this survey, as they are trained on publicly available, non-sensitive information. This work primarily aims to introduce the construction and application of LLMs in molecular biology, focusing on advancing scientific research without generating harmful content.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xuguang Ai and Ramakanth Kavuluru. 2023. End-to-end models for chemical–protein interaction extraction: Better tokenization and span-based pipeline strategies. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 610–618. IEEE.

Zhenyu Bi, Sajib Acharjee Dip, Daniel Hajialigol, Sindhura Kommu, Hanwen Liu, Meng Lu, and Xuan Wang. 2024. AI for biomedicine in the era of large language models. *arXiv preprint arXiv:2403.15673*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

Eli M Carrami and Sahand Sharifzadeh. 2024. PQA: Zero-shot protein question answering for free-form scientific enquiry with large language models. *arXiv preprint arXiv:2402.13653*.

Qiyuan Chen and Cheng Deng. 2023. Bioinfo-Bench: A simple benchmark framework for llm bioinformatics skills evaluation. *bioRxiv*.

Zhiyuan Chen, Tianhao Chen, Chenggang Xie, Yang Xue, Xiaonan Zhang, Jingbo Zhou, and Xiaomin Fang. 2024. Unifying sequences, structures, and descriptions for any-to-any protein generation with the large multimodal model HelixProtX. *arXiv preprint arXiv:2407.09274*.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Carl Edwards, Ziqing Lu, Ehsan Hajiramezanali, Tommaso Biancalani, Heng Ji, and Gabriele Scalia. 2024. Molcap-arena: A comprehensive captioning benchmark on language-enhanced molecular property prediction. *arXiv preprint arXiv:2411.00737*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024a. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Yin Fang, Kangwei Liu, Ningyu Zhang, Xinle Deng, Penghui Yang, Zhuo Chen, Xiangru Tang, Mark Gerstein, Xiaohui Fan, and Huajun Chen. 2024b. ChatCell: Facilitating single-cell analysis with natural language. *arXiv preprint arXiv:2402.08303*.

Daniel Flam-Shepherd and Alán Aspuru-Guzik. 2023. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and pdb files. *arXiv preprint arXiv:2305.05708*.

Zhangyang Gao, Chen Tan, and Stan Z Li. 2024a. FoldToken3: Fold structures worth 256 words or less. *bioRxiv*.

Zhangyang Gao, Cheng Tan, and Stan Z Li. 2023. VQPL: Vector quantized protein language. *arXiv preprint arXiv:2310.04985*.

Zhangyang Gao, Cheng Tan, and Stan Z Li. 2024b. FoldToken2: Learning compact, invariant and generative protein structure language. *bioRxiv*.

Zhangyang Gao, Cheng Tan, and Stan Z Li. 2024c. FoldToken4: Consistent & hierarchical fold language. *bioRxiv*.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. 2024d. FoldToken: Learning protein language via vector quantization and beyond. *arXiv preprint arXiv:2403.09673*.

Benoit Gaujac, Jérémie Donà, Liviu Copoiu, Timothy Atkinson, Thomas Pierrot, and Thomas D Barrett. 2024. Learning the language of protein structure. *arXiv preprint arXiv:2405.15840*.

Laura H Goetz and Nicholas J Schork. 2018. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963.

Xiao-Yu Guo, Yi-Fan Li, Yuan Liu, Xiaoyong Pan, and Hong-Bin Shen. 2024. ProtDAT: A unified framework for protein sequence design from any protein text description. *arXiv preprint arXiv:2412.04069*.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. 2024. Simulating 500 million years of evolution with a language model. *bioRxiv*.

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

Haonan He, Yuchen Ren, Yining Tang, Ziyang Xu, Junxian Li, Minghao Yang, Di Zhang, Dong Yuan, Tao Chen, Shufei Zhang, Yuqiang Li, Nanqing Dong, Wanli Ouyang, Dongzhan Zhou, and Peng Ye. 2024a. Biology instructions: A dataset and benchmark for multi-omics sequence understanding capability of large language models. *arXiv preprint arXiv:2412.19191*.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. 2024b. Lucaone: Generalized biological foundation model with unified nucleic acid and protein language. *bioRxiv*, pages 2024–05.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Mingjia Huo, Han Guo, Xingyi Cheng, Digvijay Singh, Hamidreza Rahmani, Shen Li, Philipp Gerlof, Trey Ideker, Danielle A Grotjahn, Elizabeth Villa, et al. 2024. Multi-modal large language model enables protein function prediction. *bioRxiv*.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2024. Chain-of-thoughts for molecular understanding. *arXiv preprint arXiv:2410.05610*.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. ProLLM: Protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *bioRxiv*, pages 2024–04.

Wei Lan, Guohang He, Mingyang Liu, Qingfeng Chen, Junyue Cao, and Wei Peng. 2024. Transformer-based single-cell language model: A survey. *arXiv preprint arXiv:2407.13205*.

Hongliang Li, Yihe Pang, and Bin Liu. 2021. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. *Nucleic Acids Research*, 49:e129 – e129.

Jiatong Li, Junxian Li, Yunqing Liu, Dongzhan Zhou, and Qing Li. 2024a. Tomg-bench: Evaluating llms on text-based open molecule generation. *arXiv preprint arXiv:2412.14642*.

Jiatong Li, Wei Liu, Zhihao Ding, Wenqi Fan, Yuqiang Li, and Qing Li. 2024b. Large language models are in-context molecule learners. *arXiv preprint arXiv:2403.04197*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Pan Li, Xingyi Cheng, Le Song, and Eric Xing. 2024c. Retrieval augmented protein language models for protein structure prediction. *bioRxiv*.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024d. Towards 3D molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, Zicheng Ma, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao, and Jun Zhang. 2023a. Tokenizing foldable protein structures with machine-learned artificial amino-acid vocabulary. *bioRxiv*.

Xinyi Lin, Gelei Deng, Yuekang Li, Jingquan Ge, Joshua Wing Kei Ho, and Yi Liu. 2024. GeneRAG: Enhancing large language models with gene-related task by retrieval-augmented generation. *bioRxiv*.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023b. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Andrew Liu, Axel Elaldi, Nathan Russell, and Olivia Viessmann. 2024b. Bio2Token: All-atom tokenization of any biomolecular structure with mamba. *arXiv preprint arXiv:2410.19110*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024d. Git-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638, Singapore. Association for Computational Linguistics.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024e. ProtT3: Protein-to-text generation for text-based protein understanding. *arXiv preprint arXiv:2405.12564*.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, et al. 2024. nach0: Multimodal natural and chemical languages foundation model. *Chemical Science*, 15(22):8380–8389.

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. *arXiv preprint arXiv:2403.08192*.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. ProLLaMA: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Chang Ma, Haiteng Zhao, Lin Zheng, Jiayi Xin, Qintong Li, Lijun Wu, Zhihong Deng, Yang Young Lu, Qi Liu, Sheng Wang, and Lingpeng Kong. 2024. Retrieved sequence augmentation for protein representation learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1738–1767, Miami, Florida, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shuto Negishi, James H Girsch, Elizabeth L Siegler, Evandro D Bezerra, Kotaro Miyao, and R Leo Sakemura. 2024. Treatment strategies for relapse after CAR T-cell therapy in B cell lymphoma. *Frontiers in Pediatrics*, 11:1305657.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. 2024a. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

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

Joseph Nowell, Eleanor Blunt, and Paul Edison. 2023. Incretin and insulin signaling as novel therapeutic targets for alzheimer's and parkinson's disease. *Molecular Psychiatry*, 28(1):217–229.

Fang-Shu Ou, Stefan Michiels, Yu Shyr, Alex A Adjei, and Ann L Oberg. 2021. Biomarker discovery and validation: statistical considerations. *Journal of Thoracic Oncology*, 16(4):537–545.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, and Rui Yan. 2024. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

Rafael Josip Penić, Tin Vlašić, Roland G. Huber, Yue Wan, and Mile Šikić. 2024. RiNALMo: General-purpose RNA language models can generalize well on structure prediction tasks. *arXiv preprint arXiv:2403.00043*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nikša Praljak, Hugh Yeh, Miranda Moore, Michael Socolich, Rama Ranganathan, and Andrew L Ferguson. 2024. Natural language prompts guide the design of novel functional protein sequences. *bioRxiv*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie P Lopez, Alexander Laterre, Maren Lang, et al. 2024. ChatNT: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*.

Anna C Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, et al. 2024. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*.

Patricia Sebastian-Leon, Enrique Vidal, Pablo Minguez, Ana Conesa, Sonia Tarazona, Alicia Amadoz, Carmen Armero, Francisco Salavert, Antonio Vidal-Puig, David Montaner, et al. 2014. Understanding disease mechanisms with models of signaling pathway activities. *BMC systems biology*, 8:1–19.

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

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. 2025. CAT: Content-adaptive image tokenization. *arXiv preprint arXiv:2501.03120*.

Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. 2024. TourSynbio: A multimodal large model and agent framework to bridge text and protein sequences for protein engineering. *arXiv preprint arXiv:2408.15299*.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2023. SciEval: A multi-level large language model evaluation benchmark for scientific research. *arXiv preprint arXiv:2308.13149*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Isabela Victorino da Silva Amatto, Nathalia Gonsales da Rosa-Garzon, Flavio Antonio de Oliveira Simoes, Fernanda Santiago, Nathalia Pereira da Silva Leite, Julia Raspante Martins, and Hamilton Cabral. 2022. Enzyme engineering and its industrial applications. *Biotechnology and Applied Biochemistry*, 69(2):389–409.

Mai Ha Vu, Rahmad Akbar, Philippe A. Robert, Bartlomiej Swiatczak, Geir Kjetil Ferkingstad Sandve, Victor Greiff, and Dag Trygve Tryslew Haug. 2022. Linguistically inspired roadmap for building biologically reliable protein language models. *Nature Machine Intelligence*, pages 1–12.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. 2024a. ProtChatGPT: Towards understanding proteins with large language models. *arXiv preprint arXiv:2402.09649*.

Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024b. BioRAG: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023. InstructProtein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*.

Wei Wu, Chao Wang, Liyi Chen, Mingze Yin, Yiheng Zhu, Kun Fu, Jieping Ye, Hui Xiong, and Zheng Wang. 2024. Structure-enhanced protein instruction tuning: Towards general-purpose protein understanding. *arXiv preprint arXiv:2410.03553*.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, et al. 2025. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv preprint arXiv:2502.07527*.

Hongwang Xiao, Wenjun Lin, Hui Wang, Zheng Liu, and Qiwei Ye. 2024a. OPI: An open instruction dataset for adapting large language models to protein-related tasks. In *NeurIPS 2024 Workshop: Foundation Models for Science: Progress, Opportunities, and Challenges*.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. 2024b. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*.

Binghao Yan, Yunbi Nam, Lingyao Li, Rebecca A Deek, Hongzhe Li, and Siyuan Ma. 2024a. Recent advances in deep learning and language models for studying the microbiome. *arXiv preprint arXiv:2409.10579*.

Wilson Yan, Matei Zaharia, Volodymyr Mnih, Pieter Abbeel, Aleksandra Faust, and Hao Liu. 2024b. Elastictok: Adaptive tokenization for image and video. *arXiv preprint arXiv:2410.08368*.

Mehdi Yazdani-Jahromi, Mangal Prakash, Tommaso Mansi, Artem Moskalev, and Rui Liao. 2024. HELM: Hierarchical encoding for mRNA language modeling. *arXiv preprint arXiv:2410.12459*.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. 2024. ProteinBench: A holistic evaluation of protein foundation models. *arXiv preprint arXiv:2409.06744*.

Kamyar Zeinalipour, Neda Jamshidi, Monica Bianchini, Marco Maggini, and Marco Gori. 2024. Design proteins using large language models: Enhancements and comparative analyses. In *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pages 34–47, Bangkok, Thailand. Association for Computational Linguistics.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024a. ChemLLM: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Jianrong Zhang, Hehe Fan, and Yi Yang. 2024b. Protein captioning: Bridging the gap between protein sequences and natural languages. *ACM Trans. Multimedia Comput. Commun. Appl.*

Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, Yizhen Luo, and Zaiqing Nie. 2024a. LangCell: Language-cell pre-training for cell identity understanding. *arXiv preprint arXiv:2405.06708*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. 2024b. ChemDFM: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024. Benchmarking large language models for molecule prediction tasks. *arXiv preprint arXiv:2403.05075*.

Xibin Zhou, Chenchen Han, Yingqi Zhang, Jin Su, Kai Zhuang, Shiyu Jiang, Zichen Yuan, Wei Zheng, Fengyuan Dai, Yuyang Zhou, Yuyang Tao, Dan Wu, and Fajie Yuan. 2025a. Decoding the molecular language of proteins with evola. *bioRxiv*.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

Zhihan Zhou, Robert Riley, Satria Kautsar, Weimin Wu, Rob Egan, Steven Hofmeyr, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Fengchen Liu, et al. 2025b. Genomeocean: An efficient genome foundation model trained on large-scale metagenomic assemblies. *bioRxiv*, pages 2025–01.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, et al. 2024. InstructBioMol: Advancing biomolecule understanding and design following human instructions. *arXiv preprint arXiv:2410.07919*.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. 2024. ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8950–8963, Bangkok, Thailand. Association for Computational Linguistics.

A Differences and similarities of natural and biological languages

Natural and biological languages both encode complex, structured information for communication, but differ fundamentally in meaning derivation—natural language relies on linguistic and contextual relationships, while biological language encodes functions through structural properties and biomolecular interactions.

B Biological language models

Single-cell language models. Single-cell sequencing enables the study of individual cells, allowing researchers to analyze cellular heterogeneity and unique characteristics at high resolution. Single-cell language models, such as scGPT (Cui et al., 2024) and Nicheformer (Schaar et al., 2024), are

Table 5: Natural Language *v.s.* Biological Language: Differences and Similarities

	Feature	Natural Language	Biological Language
Differences	**Origin**	Human-developed communication	Naturally evolved molecular codes
	Symbols	Words, phonemes, sentences	Nucleotides, amino acids, molecules
	Structure	Grammar and syntax rules	Biochemical and structural constraints
	Meaning	Context-dependent interpretation	Functional molecular outcomes
	Ambiguity	Common and resolved through context	Minimally ambiguous, highly structured
	Evolution	Cultural and historical evolution	Evolution via mutation and selection
Similarities	**Symbolic Systems**	Words, sentences, and texts form communication structures	DNA, RNA, and protein sequences encode biological information
	Hierarchical Structure	Follows syntactic and semantic structures (words → phrases → sentences)	Organized from genome to proteins, cellular signaling, and systems biology
	Context Dependency	Word meanings depend on surrounding text (e.g., "bank" may refer to finance or a riverbank)	The function of a sequence depends on its context and environment
	Redundancy and Robustness	Contains synonyms, ambiguities, and grammatical flexibility	Genetic code has redundancy (e.g., synonymous codons) and error tolerance mechanisms

computational models designed to process and interpret single-cell sequencing data. These models facilitate the exploration of genomic, transcriptomic, proteomic, and epigenomic information at the single-cell level, advancing our understanding of cellular function and differentiation (Lan et al., 2024).

Protein language models. Proteins are complex macromolecules composed of amino acid chains that play essential roles in biological processes. Protein language models apply linguistical ideas to model protein sequences, facilitating function prediction, structure prediction, and sequence design. These models have garnered significant attention due to their potential in protein biology (Vu et al., 2022).

RNA language models. RNA is an important biomolecule involved in genetic information transfer and protein synthesis. RNA language models are specialized computational models designed to process and analyze RNA sequences (Penić et al., 2024). They enable the prediction of structural and functional attributes, such as secondary structure base-pairing probabilities and solvent accessibility, which are crucial for understanding RNA function and binding mechanism (Bi et al., 2024).

DNA language models. DNA encodes genetic instructions essential for the growth, development, and reproduction of all living organisms. DNA language models treat DNA sequences as structured linguistic data, enabling pattern recognition and function prediction (Yan et al., 2024a). These models facilitate the study of genome architecture, regulatory elements, and evolutionary dynamics, advancing genomics research and precision medicine.

Small molecule language models. Small molecule language models are specifically designed to analyze and predict the chemical properties and interactions of small molecules (Flam-Shepherd and Aspuru-Guzik, 2023). By treating small molecules, such as drugs, metabolites, and other low molecular weight compounds, as sequences or graphs, these models capture their unique characteristics, contributing to advancements in drug discovery and cheminformatics.