# Biological Sequence with Language Model Prompting: A Survey

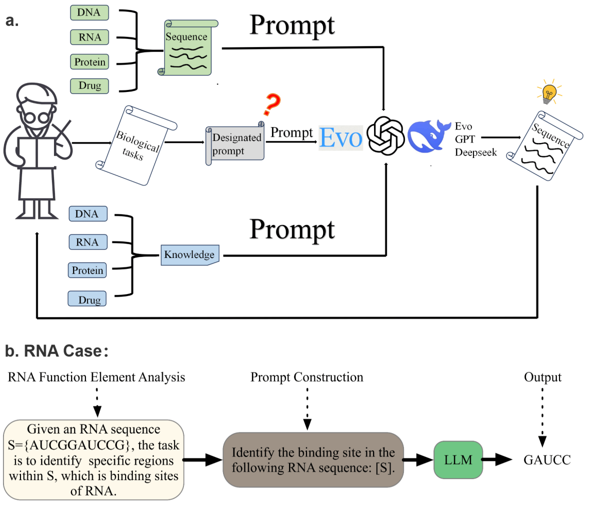
Anonymous ACL submission

## Abstract

Large Language models (LLMs) have emerged as powerful tools for addressing challenges across diverse domains. Notably, recent studies have demonstrated that LLMs can substantially improve the efficiency of biomolecular analysis and synthesis, garnering increasing attentions across both academic research and medical applications. In this paper, we systematically investigate how LLMs, guided by prompt-based methodologies,can be applied to biological sequence analysis, including DNA, RNA, proteins, and tasks related to drug discovery. Specifically, we explore how prompt engineering enables LLMs to tackle domain-specific problems, such as promoter sequence prediction, protein structure modeling, and drug-target binding affinity prediction, often in scenarios with limited labeled data. Furthermore, our discussion highlights the transformative potential of prompting in bioinformatics while addressing key challenges such as data scarcity, multimodal fusion, and computational resource limitations. This paper is intended to serve both as a foundational resource for newcomers and as a springboard for ongoing innovation in this rapidly evolving field of study.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable advancements, primarily due to their capabilities in modeling the hidden relationships within textual sequences (Achiam et al., 2023; Dubey et al., 2024). This innovation presents a compelling opportunity for bioinformatics, where biological sequences (e.g., DNA, RNA, and proteins) exhibit structural and statistical similarities to natural languages (Searls, 1997). By leveraging LLMs, researchers can uncover meaningful patterns from these sequences, leading to notable breakthroughs in diverse downstream tasks, such as classification, structure prediction and drug discovery, as illustrated in the general workflow of Figure 1**a**. Figure 1**b** further highlights a concrete example in the context of RNA function analysis, which underscores the importance of effective prompt engineering in enabling LLMs to perform meaningful biological inference even with limited labeled data. However, despite this potential, several challenges remain in applying LLMs to biological sequence analysis. One of the most pressing issues is data scarcity, as labeled biological datasets are always expensive and labor-intensive to obtain. This limitation significantly restricts the effectiveness of traditional supervised learning approaches.



Given an RNA sequence S={AUCGGAUCCG}, the task is to identify specific regions within S, which is binding sites of RNA.

Identify the binding site in the following RNA sequence: [S].

Figure 1: Diagram of LLM prompting for de novo design of biomolecules. (**a**) General workflow illustrating how biological sequences and task-specific knowledge are used to construct prompts. These prompts are then processed by LLMs to generate biologically coherent and task-aligned sequence predictions. (**b**) Diagram of an RNA-specific case, where the task is to identify functional binding sites within a given RNA sequence. A prompt is constructed accordingly and used to guide the LLM in generating the desired output sequence.

To address this issue, prompt engineering has emerged as a powerful strategy to enhance the adaptability of LLMs in biological contexts (Zaghir et al., 2024; Zhou and Ngo, 2024; Chang et al., 2025). Specifically, prompt-based methods leverage in-context learning, allowing LLMs to perform zero-shot and few-shot learning on biological tasks with minimal labeled data. Representative models (e.g., BERT (Devlin, 2018), GPT (Achiam et al., 2023), ProtBERT (Brandes et al., 2022) and Evolutionary Scale Modeling (ESM) (Lin et al., 2022)) have been successfully adapted to biological sequences through precisely designed prompts, allowing them to generalize across a range of tasks represented by promoter sequence prediction, protein structure modeling, and drug-target binding affinity prediction. Figure 2 showcases this rapid evolution, highlighting key milestones from foundational models to recent prompt-driven advances.

**Organization of This Survey:** In this paper, we conduct the first survey of recent advancements in biological sequence analysis through language model prompting. We begin by introducing the fundamentals of biological sequences and explaining how prompt engineering facilitates LLMs applications in various downstream taks accross different domains (§2). Representative examples of these applications are visually summarized in Figure 3, which categorizes prompt-based workflows across four major biological areas, highlighting the role of prompt construction in guiding LLMs outputs. We then present a detailed survey of prompting methodologies (§3), classifying existing approaches by biological domains and task-specific objectives. A structured taxonomy of involved literature is provided in Figure 4. In parallel, we examine the transformative role of AlphaFold (Jumper et al., 2021; Abramson et al., 2024) and the ESM series (Rives et al., 2019; Lin et al., 2022, 2023) (§3), highlighting their contributions to protein structure prediction and proteome-scale modeling. Next, we outline several key challenges (§4) such as data scarcity and high labeling costs then explore future research directions (§5), focusing on multi-modal prompt fusion, efficient adaptation techniques and data-centric annotation strategies. Finally, we conclude this survey (§6) by summarizing key insights and underscoring the role of prompt engineering in advancing AI-driven biological research.

## 2 Biological Sequence Prompting Tasks

In this section, we first introduce the concept of biological sequences and how language modeling and prompt engineering fit into bioinformatics. Then, we outline how various biological tasks can be formulated as natural language processing (NLP) problems and provide a concise categorization of their application domains (DNA, RNA, proteins, and drug discovery), setting the stage for the methodological discussions in subsequent sections.

### 2.1 Biological Sequences

Biological sequences, such as DNA, RNA, and proteins, can be viewed as linear arrangements of tokens from their respective alphabets. For instance, DNA and RNA use nucleotides $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T/U}\}$, while proteins typically involve 20 standard amino acids. Formally, a biological sequence of length $L$ is denoted as

$$S = \{x_1, \ldots, x_L\}, \tag{1}$$

where each token $x_i$ is drawn from an alphabet $\mathcal{A}$. These sequences carry crucial information for cellular processes, such as gene expression, protein folding, and molecular interactions. Understanding and modeling these sequences is central to many tasks in computational biology.

### 2.2 Prompt Engineering

Prompts further enhance the power of LLMs by integrating task-specific cues into the input of the model. This is achieved through textual templates or continuous embeddings, which we denote abstractly as:

$$T = f(S, P; \theta), \tag{2}$$

where $S$ represents biological sequence, $P$ encodes domain-specific information as a prompt, and $\theta$ denotes the model parameters. Prompting is particularly valuable in data-scarce scenarios (e.g., zero-shot or few-shot learning), guiding the model to focus on relevant biological patterns.

### 2.3 Overview of Prompt-Based Task Mapping

Biological tasks such as promoter prediction or modeling of protein-ligand binding interactions can be effectively reinterpreted as NLP tasks using prompts. For instance, in promoter prediction, specific segments of a DNA sequence can be replaced with [MASK] tokens, transforming the task into a masked language modeling problem. Similarly, protein-ligand binding affinity prediction can be framed as a question-answering or fill-in-the-blank prompt, focusing on molecular compatibility. By carefully designing prompts, researchers leverage the extensive pretraining of LLMs to achieve strong performance even with limited labeled data.

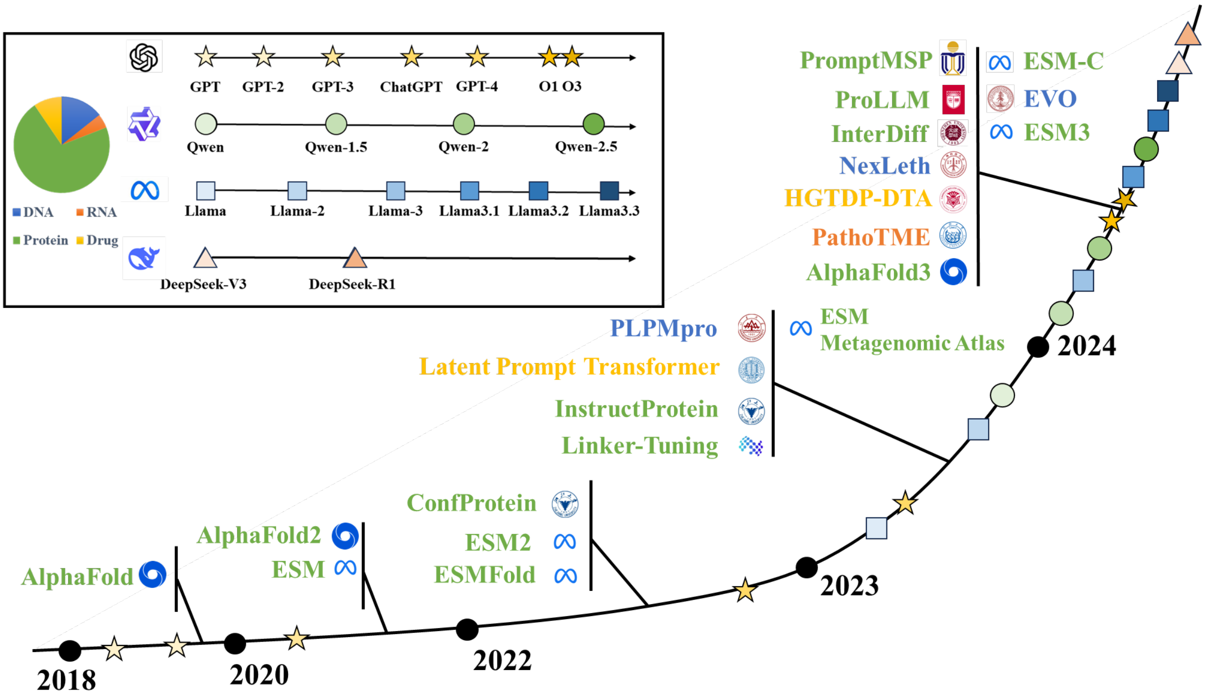


Figure 2: Milestones in the evolution of LLMs and prompt-based advances for computational biology.

In this section, we provide an overview of four key application areas in biological sequence analysis:

### 2.3.1 DNA

**Promoter Identification.** Detects promoter regions in DNA sequences, capturing key motifs such as the TATA-box to understand gene regulation. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$H = M_{\text{DNA}}(S, P_{\text{prom}}), \tag{3}$$

where $S$ is the input DNA sequence, $M_{\text{DNA}}$ is the large language model of DNA for promoter identification, $P_{\text{prom}}$ is is the constructed prompt, and $H$ is the predicted promoter region with functional motifs.

For example, given a sequence $S = \{ATGCGATACTAGGATATAAGCTAG\}$. To detect the promoter region in this DNA sequence, we design this prompt: "Locate the promoter region between positions X-Y in [S] containing TATA-box motifs." Then the DNA sequence and constructed prompts are inputted into the language models to generate more accurate results.

**Mechanism Explanation.** Generates interpretable insights into synthetic lethality (SL) mechanisms, aiding in identifying potential cancer drug targets. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$J = M_{\text{DNA}}(S_g, P_{\text{sl}}), \tag{4}$$

where $S_g$ is the input gene pair, $M_{\text{DNA}}$ is the large language model of DNA for mechanism explanation, $P_{\text{sl}}$ is the constructed prompt, and $J$ is the structured explanation of SL interactions.

For instance, given a gene pair $S_g = \{BRCA1, PARP1\}$. To explain the SL mechanism, we construct a corresponding prompt: "Explain the synthetic lethality mechanism between [Gene A] and [Gene B], focusing on DNA repair pathways". Then, the gene pair and constructed prompts are inputted into the language models to generate more accurate results.

### 2.3.2 RNA

**RNA Functional Element Analysis.** Identifies and characterizes splicing signals, regulatory elements, and sequence motifs associated with gene expression regulation. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$R = M_{\text{RNA}}(S, P_f), \tag{5}$$

where $S$ is the input RNA sequence, $M_{\text{RNA}}$ is the large language model of RNA for functional element analysis, $P_f$ is the constructed prompt, and $R$ is the generated response, which corresponds to the functional element $E$.

Moreover, we take a specific example to illustrate how to use a prompt-based method to model RNA-related tasks as NLP tasks. For instance, given an RNA sequence $S = \{AUCGGAUCCG\}$, we want to identify specific regions within $S$, which are binding sites of RNA. We can construct a corresponding prompt: "Identify the binding site in the following RNA sequence: [$S$]". Then, the RNA sequence and constructed prompts are inputted into the language models to generate more accurate results (Figure 1**b**).

**Cell Type Annotation.** Automates the classification of cell types in single-cell RNA sequencing (scRNA-seq) data, improving accuracy in diverse and noisy datasets. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$L = M_{\text{RNA}}(S_c, P_{\text{cell}}), \tag{6}$$

where $S_c$ is the input gene expression profile, $M_{\text{RNA}}$ is large language model of RNA for cell type annotation, $P_{\text{cell}}$ is the constructed prompt, and $L$ corresponds to the predicted cell type labels with confidence scores.

Moreover, we take a specific example to illustrate how to use a prompt-based method to model RNA-related tasks as NLP tasks. For instance, given a cell expression profile $S_c = \{CD3E : 12.8, CD8A : 9.4, CD19 : 0.3, MS4A1 : 0.1\}$, we construct a corresponding prompt: "Classify this scRNA-seq cell's type based on top expressed genes: [S], and include canonical markers". Then, the cell expression profile and constructed prompts are inputted into the language models to generate more accurate results.

### 2.3.3 Protein

**Protein Structure Modeling and Prediction.** Focuses on determining the three-dimensional structure of proteins from amino acid sequences, which is crucial for understanding function and interactions. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$\Phi = M_{\text{protein}}(S, P_{\text{fold}}) \tag{7}$$

where $S$ is the input amino acid sequence, $M_{\text{protein}}$ is large language model of protein for structure modeling and prediction, $P_{\text{fold}}$ is the is the constructed prompt, and $\Phi$ corresponds to the predicted structural coordinates.

For example, given a sequence $S = \{GVNPGVAPLSLLI\}$, we can construct a corresponding prompt: "Predict the tertiary structure topology for [S] with secondary structure annotations". Then, the protein sequence and constructed prompts are inputted into the language models to generate more accurate results.

**Molecular Interaction Modeling.** Studies how proteins interact with other molecules, including ligands and other proteins, to inform drug design and functional analysis. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task through constructing proper prompts, which can be formally defined as follows:

$$\Gamma = M_{\text{protein}}(S_l, P_{\text{interact}}), \tag{8}$$

where $S_l$ is the input molecular pair, $M_{\text{protein}}$ is large language model of protein for molecular interaction modeling, $P_{\text{interact}}$ is the is the constructed prompt, and $\Gamma$ corresponds to the predicted binding parameters.

For instance, given a kinase-ligand pair $S_l = \{EGFRkinase : MGPSV..., Gefitinib : C1 = CN = CC = C1\}$, we can construct a corresponding prompt: "Predict binding mode between EGFR kinase and Gefitinib, identifying critical hydrogen bonds and hydrophobic contacts". Then, the molecular pair and constructed prompts are inputted into the language models to generate more accurate results.

**Protein Language-Based Generation.** Explores the generation of protein sequences and functional annotations using language-inspired models, facilitating de novo protein design. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$\Psi = M_{\text{protein}}(P_{\text{design}}, \theta), \tag{9}$$

where $\theta$ is functional constraints, $M_{\text{protein}}$ is a large language model of protein for de novo protein design, $P_{\text{design}}$ is the constructed prompt, and $\Psi$ corresponds to the generated protein sequence with structural and functional metadata.

For example, given a target,to engineer a thermostable enzyme, we can construct a corresponding prompt: "Generate a $\beta$-lactamase variant with enhanced thermal stability and maintained catalytic efficiency". Then, the constructed prompts is inputted into the language models to generate more accurate results.

**Other Protein-Related Tasks.** Includes *Polypeptide Design*, *Conformation Perception*, and *Protein Interaction Reasoning*, expanding applications in structural and functional biology. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$K = M_{\text{protein}}(S_r, P_{\text{task}}), \tag{10}$$

where $S_r$ are the task-specific inputs, $M_{\text{protein}}$ is a large language model of protein for other tasks, $P_{\text{task}}$ is the constructed prompt, and $K$ corresponds to multimodal outputs spanning sequences, structures, and mechanistic insights.

For easy understanding, we will illustrate how these protein-related tasks can be modeled as NLP tasks using a cue-based approach by following three concrete examples. First, in order to generate antimicrobial peptides, we can construct a corresponding prompt: "Design a 15-residue cationic $\alpha$-helical peptide targeting Gram-negative bacteria with <10% hemolysis". Second, given a kinase sequence $S_r = \{IGPGRAFVT\}$, in order to predict conformational, we can construct a corresponding prompt: "Predict conformational changes upon ATP binding". Finally, given a ubiquitin-ligase pair $S_r = \{Ubiquitin, E6AP\}$. To reason about protein interactions, we can construct a corresponding prompt: "Infer recognition mechanism for Ub-E6AP complex formation". Then, the inputs and constructed prompts are inputted into the language models to generate more accurate results.

#### 2.3.4 Drug Discovery

**Drug-Target Binding Prediction.** Focuses on estimating the binding affinity between drugs and their molecular targets, aiding in the identification of potential therapeutics. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$N = M_{\text{drug}}(S_d, S_t, P_{\text{bind}}), \tag{11}$$

where $S_d$ is the drug molecular structure, $S_t$ is the target protein sequence, $M_{\text{drug}}$ is the large language model of drug for target binding prediction, $P_{\text{bind}}$ is the constructed prompt, and N corresponds to predicted binding metrics.

For example, given the anticancer drug Gefitinib and EGFR kinase, we can construct a corresponding prompt: "Predict binding affinity and critical interactions between Gefitinib and EGFR kinase".Then, the inputs and constructed prompts are inputted into the language models to generate more accurate results.

**Molecular Design.** Involves generating and optimizing molecular structures to achieve desired pharmaceutical properties, such as bioavailability and synthetic accessibility. Benefiting from the remarkable performance of prompting, we effectively reinterpreted this task as an NLP task by constructing proper prompts, which can be formally defined as follows:

$$U = M_{\text{drug}}(\theta, P_{\text{design}}), \tag{12}$$

where $\theta$ is target property specifications, $M_{\text{drug}}$ is a large language model of drug for molecular design, $P_{\text{design}}$ is the constructed prompt, and U corresponds to the generated molecule.

For instance, to design a non-steroidal anti-inflammatory drug (NSAID) with reduced gastrointestinal toxicity, we can construct a corresponding prompt: "Generate a COX-2 selective inhibitor with $IC_{50} < 10nM$, LogP 2.5–3.5, and $> 80\%$ plasma stability at 24h". Then, the constructed prompts are inputted into the language models to generate more accurate results.

In the next sections, we will delve deeper into the specific methods, discussing their performance and how prompt-based strategies are reshaping biological sequence analysis. By encoding domain knowledge through carefully designed prompts, LLMs have become powerful tools for tackling complex tasks in DNA, RNA, protein, and drug discovery.

## 3 Prompting Applications in Biological Sequences

Prompting technology, which is especially suitable for zero/few sample learning by designing prompts for pre-trained models to guide them to accomplish

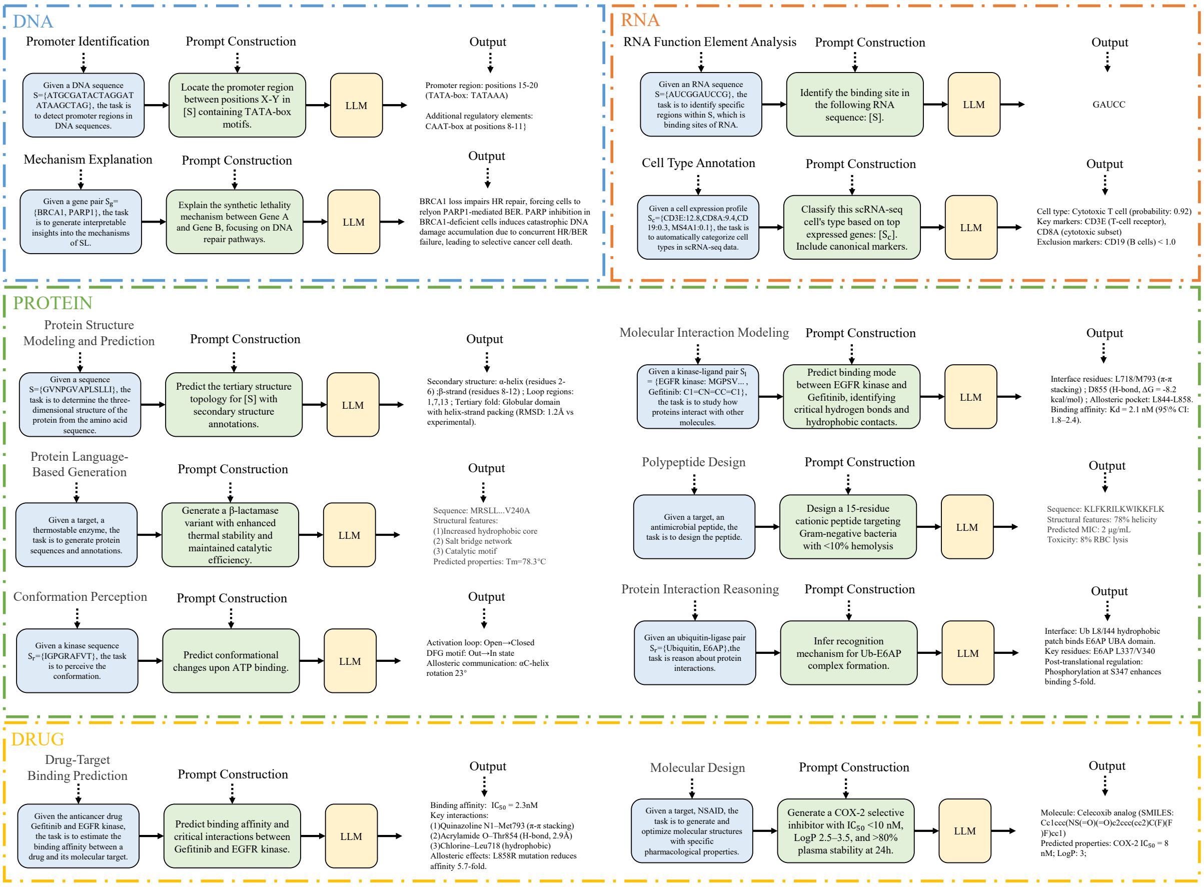

DNA
Promoter Identification
Prompt Construction
Output
Given a DNA sequence S={ATGCGATACTAGGAT ATAAGCTAG}, the task is to detect promoter regions in DNA sequences.
Locate the promoter region between positions X-Y in [S] containing TATA-box motifs.
LLM
Promoter region: positions 15-20 (TATA-box: TATAAA)
Additional regulatory elements: CAAT-box at positions 8-11}
Mechanism Explanation
Prompt Construction
Output
Given a gene pair $S_g$={BRCA1, PARP1}, the task is to generate interpretable insights into the mechanisms of SL.
Explain the synthetic lethality mechanism between Gene A and Gene B, focusing on DNA repair pathways.
LLM
BRCA1 loss impairs HR repair, forcing cells to relyon PARP1-mediated BER. PARP inhibition in BRCA1-deficient cells induces catastrophic DNA damage accumulation due to concurrent HR/BER failure, leading to selective cancer cell death.
RNA
RNA Function Element Analysis
Prompt Construction
Output
Given an RNA sequence S={AUCGGAUCCG}, the task is to identify specific regions within S, which is binding sites of RNA.
Identify the binding site in the following RNA sequence: [S].
LLM
GAUCC
Cell Type Annotation
Prompt Construction
Output
Given a cell expression profile $S_c$={CD3E:12.8,CD8A:9.4,CD 19:0.3, MS4A1:0.1}, the task is to automatically categorize cell types in scRNA-seq data.
Classify this scRNA-seq cell's type based on top expressed genes: [$S_c$]. Include canonical markers.
LLM
Cell type: Cytotoxic T cell (probability: 0.92)
Key markers: CD3E (T-cell receptor), CD8A (cytotoxic subset)
Exclusion markers: CD19 (B cells) < 1.0
PROTEIN
Protein Structure Modeling and Prediction
Prompt Construction
Output
Given a sequence S={GVNPGVAPLSLLI}, the task is to determine the three-dimensional structure of the protein from the amino acid sequence.
Predict the tertiary structure topology for [S] with secondary structure annotations.
LLM
Secondary structure: α-helix (residues 2-6) ;β-strand (residues 8-12) ; Loop regions: 1,7,13 ; Tertiary fold: Globular domain with helix-strand packing (RMSD: 1.2Å vs experimental).
Molecular Interaction Modeling
Prompt Construction
Output
Given a kinase-ligand pair $S_l$ = {EGFR kinase: MGPSV... , Gefitinib: C1=CN=CC=C1}, the task is to study how proteins interact with other molecules.
Predict binding mode between EGFR kinase and Gefitinib, identifying critical hydrogen bonds and hydrophobic contacts.
LLM
Interface residues: L718/M793 (π-π stacking) ; D855 (H-bond, ΔG = -8.2 kcal/mol) ; Allosteric pocket: L844-L858. Binding affinity: Kd = 2.1 nM (95\% CI: 1.8–2.4).
Protein Language-Based Generation
Prompt Construction
Output
Given a target, a thermostable enzyme, the task is to generate protein sequences and annotations.
Generate a β-lactamase variant with enhanced thermal stability and maintained catalytic efficiency.
LLM
Sequence: MRSLL...V240A
Structural features:
(1)Increased hydrophobic core
(2) Salt bridge network
(3) Catalytic motif
Predicted properties: Tm=78.3°C
Polypeptide Design
Prompt Construction
Output
Given a target, an antimicrobial peptide, the task is to design the peptide.
Design a 15-residue cationic peptide targeting Gram-negative bacteria with <10% hemolysis
LLM
Sequence: KLFKRILKWIKKFLK
Structural features: 78% helicity
Predicted MIC: 2 μg/mL
Toxicity: 8% RBC lysis
Conformation Perception
Prompt Construction
Output
Given a kinase sequence $S_r$={IGPGRAFVT}, the task is to perceive the conformation.
Predict conformational changes upon ATP binding.
LLM
Activation loop: Open→Closed
DFG motif: Out→In state
Allosteric communication: αC-helix rotation 23°
Protein Interaction Reasoning
Prompt Construction
Output
Given an ubiquitin-ligase pair $S_r$={Ubiquitin, E6AP},the task is reason about protein interactions.
Infer recognition mechanism for Ub-E6AP complex formation.
LLM
Interface: Ub L8/I44 hydrophobic patch binds E6AP UBA domain.
Key residues: E6AP L337/V340
Post-translational regulation: Phosphorylation at S347 enhances binding 5-fold.
DRUG
Drug-Target Binding Prediction
Prompt Construction
Output
Given the anticancer drug Gefitinib and EGFR kinase, the task is to estimate the binding affinity between a drug and its molecular target.
Predict binding affinity and critical interactions between Gefitinib and EGFR kinase.
LLM
Binding affinity: $IC_{50}$ = 2.3nM
Key interactions:
(1)Quinazoline N1–Met793 (π-π stacking)
(2)Acrylamide O–Thr854 (H-bond, 2.9Å)
(3)Chlorine–Leu718 (hydrophobic)
Allosteric effects: L858R mutation reduces affinity 5.7-fold.
Molecular Design
Prompt Construction
Output
Given a target, NSAID, the task is to generate and optimize molecular structures with specific pharmacological properties.
Generate a COX-2 selective inhibitor with $IC_{50}$ <10 nM, LogP 2.5–3.5, and >80% plasma stability at 24h.
LLM
Molecule: Celecoxib analog (SMILES: Cc1ccc(NS(=O)(=O)c2ccc(cc2)C(F)(F)F)cc1)
Predicted properties: COX-2 $IC_{50}$ = 8 nM; LogP: 3;

Figure 3: Representative prompting workflows across DNA, RNA, protein, and drug-related tasks.

specific tasks, has been widely applied across various domains. It is extensively used in natural language processing (e.g., GPT (Achiam et al., 2023), Llama (Touvron et al., 2023) for text categorization, summarization, Q&A, etc.) and bioinformatics (e.g., DNABERT (Ji et al., 2021) for recognizing DNA sequences, GPT-4 for annotating RNA-seq data (Hou and Ji, 2024), and sequential prompts for improving protein structure prediction and molecular design). By embedding task-relevant information in prompts, this technique focuses the model's attention on producing high-quality categorization, prediction, or generative output, making it an effective tool for solving complex problems due to its flexibility and data efficiency. The following sections categorize and summarize prompting methods across DNA, RNA, protein, and drug discovery domains, with a detailed taxonomy illustrated in Figure 4.

### 3.1 DNA Sequences

**NexLeth** (Zhang et al., 2024a) is a novel approach for generating natural language explanations of SL mechanisms, which are critical for cancer drug discovery. The NexLeth pipeline integrates SL knowledge graphs with personalized prompt templates, enhancing explainability and textual coherence in SL predictions.

**PLPMpro** (Li et al., 2023) is a prompt-learning framework that leverages pre-trained models such as DNABERT (Ji et al., 2021) for promoter sequence prediction. By employing soft templates and verbalizers, PLPMpro effectively captures biologically meaningful sequence patterns, such as the TATA-box motif, achieving state-of-the-art (SOTA) performance.

### 3.2 RNA Sequences

**PathoTME** (Meng et al., 2024) is a genomics-guided deep learning framework for tumor microenvironment (TME) subtype prediction. Utilizing visual prompt tuning (VPT) and domain adversarial networks, PathoTME achieves superior classification performance while addressing tissue heterogeneity challenges.

**GPTCellType** (Hou and Ji, 2024) applies GPT-4 for automated cell type annotation in single-cell RNA sequencing (scRNA-seq). GPTCellType outperforms traditional methods, demonstrating robustness to noisy datasets and underscoring the potential of LLMs for RNA-related analyses.

### 3.3 Protein Sequences

**InterDiff** (Wu et al., 2024) is a diffusion-based molecular generative model. By incorporating interaction prompts, InterDiff guides molecular design for protein-ligand interactions and achieves superior performance in predicting binding affinity and interaction specificity.

**Linker-Tuning** (Zou et al., 2023) is a lightweight adaptation method for LLMs such as ESMFold (Lin et al., 2023). This approach improves the prediction of heterodimeric protein structures and achieves competitive accuracy while reducing computational costs.

**InstructProtein** (Wang et al., 2023) is a bidirectional framework that bridges protein and human languages. By leveraging instruction tuning, this model excels at zero-shot protein function annotation and de novo sequence design.

**PromptMSP** (Gao et al., 2024) enhances multimer structure prediction through just-in-time learning and meta-learning. PromptMSP integrates conditional PPI knowledge and improves accuracy by reorganizing multimer prediction into fixed-scale tasks. The results for PDB-M show that the method outperforms the baseline method.

**ConfProtein** (Zhang et al., 2022) enhances pre-trained protein models (PTPMs) with the help of prompt learning, integrating sequence and interaction conformational cues to capture protein conformations. From the results, ConfProtein improves PPI prediction and antibody binding while maintaining sequence correlation performance, and is effectively validated on multiple datasets.

**ProLLM** (Jin et al., 2024) leverages LLMs and the Protein Chain of Thought (ProCoT) mechanism to model direct and indirect protein-protein interactions (PPIs) as inference tasks. By integrating ProtTrans embeddings and instruction fine-tuning, it achieves SOTA performance in PPI prediction.

### 3.4 Drug Discovery

**HGTDP-DTA** (Xiao et al., 2024) is a hybrid Graph-Transformer framework with dynamic prompt generation for drug-target binding affinity prediction. This model integrates both graph-based and sequence-based representations, achieving superior performance over SOTA methods on benchmark datasets.

**Latent Prompt Transformer** (Kong et al., 2024) is a generative framework for molecule design. By incorporating latent prompts into a unified architecture, the Latent Prompt Transformer achieves SOTA performance in multi-objective molecule optimization and drug-like molecule generation.

**In-Context Learning for Drug Synergy Prediction** (Edwards et al., 2023) introduces an in-context learning strategy for predicting synergistic drug combinations. By leveraging masking techniques and graph representations, this approach enhances personalized drug synergy prediction.

### 3.5 The Significance of AlphaFold and ESM

**AlphaFold.** AlphaFold (Jumper et al., 2021; Abramson et al., 2024) has revolutionized protein structure prediction, achieving near-atomic accuracy in determining 3D structures directly from amino acid sequences. By leveraging attention-based architectures, it has made notable progress toward addressing the challenge of protein folding, thereby accelerating advancements in drug discovery and enzyme engineering.

**ESM Series.** The ESM series, including ESMFold (Lin et al., 2023) and ESM-2 (Rives et al., 2019), represents a major leap in protein language modeling by enabling high-throughput structure prediction without the need for multiple sequence alignments (MSAs). ESM-3 (Hayes et al., 2025) further expands this capability, integrating sequence, structure, and function while maintaining computational efficiency. These models are invaluable for large-scale proteome analysis, facilitating the study of poorly characterized protein families.

## 4 Challenges

**Data Scarcity and High Labeling Costs.** High-quality datasets for DNA/RNA sequences, such as NexLeth (Zhang et al., 2024a) for SL gene pairs, often require manual curation and expert review, making them expensive and limited in scale, especially for rare diseases and minority species (Graefe et al., 2025). In protein studies, experimentally determined structures and reliable annotations remain insufficient, particularly for multi-chain complexes, protein interactions, and diverse conformations. Without robust experimental validation or high-confidence labels, performance gains from prompt-based methods are often hard to reproduce (Chen et al., 2024). Similarly, in drug discovery, drug-target affinity (DTA) data is scarce and expensive, limiting generalization across small samples, species, and novel targets (Pei et al., 2023).

**Difficulties in Multimodal Feature Fusion.** Bioinformatics often involves integrating sequence, structure, image, and phenotypic data. While methods like PathoTME (Meng et al., 2024) have combined visual prompts with genomic data for tumor subtype prediction, fusing high-dimensional data, such as images, protein 3D structures, transcriptomes, and molecular graphs, within a unified prompt framework remains a significant challenge (Koh et al., 2024). This complexity intensifies in tasks like protein-ligand interactions or polymer modeling, where rich contextual and dependency information is essential, yet current models struggle to scale effectively (Cao et al., 2024).

**Computational Resource Constraints.** Large-scale pre-trained models, such as ESM-2 (Rives et al., 2019) and Evo (Zhang et al., 2024c; Nguyen et al., 2024), excel at handling long sequences and large datasets but incur high computational and training costs. While higher-order models like AlphaFold-3 (Abramson et al., 2024) and ESMFold (Lin et al., 2023) offer impressive accuracy but demand substantial hardware, limiting accessibility for smaller research groups. Additionally, prompt-based approaches often require task-specific designs, such as linker-tuning or dynamic prompt generation, to adapt to different downstream tasks or data distributions (Giray, 2023). Without efficient strategies, the cost of model iteration and optimization can become prohibitively high (Ye and Durrett, 2022).

## 5 Future directions

**Data-Centric Synthetic Annotation Methods.** To address data scarcity, future work can explore semi-supervised learning, domain adaptation, and synthetic data generation (Zha et al., 2025; Hu et al., 2024; Zhang et al., 2024b). For instance,generative models can augment limited experimental samples, while active learning frameworks can guide annotation more effectively and help prompt-based models generalize in resource-poor domains (Zhao et al., 2020).

**Multi-Modal Prompt Fusion.** Beyond sequence-level prompts, unifying structural, image, and metagenomic data, among others, is a future direction (Liu et al., 2024). Meanwhile, designing consistent cross-modal prompts and specialized attention layers helps models capture more complex correlations (Ampazis and Sakketou, 2024). To illustrate current capabilities, Figure 5 presents a case study where we apply four LLMs to generate DNA sequences in response to a unified prompt requesting TD-related sequences with controlled GC content and codon usage, based on a reference sequence: GATAGAGAGACAAAGAGGAAAAGAGAGCGAGGTAGAAAACGGATACTGCCTATGCCTACTCCATCCCTCT. AlphaFold3 (Abramson et al., 2024) is then used to predict the structures from each LLM-generated DNA sequence. Structural alignment against the ground truth reveals notable variation in accuracy, with LLMs like Qwen-2.5-Max (Yang et al., 2024) achieving lower root mean square deviation (RMSD) and higher local distance difference test (LDDT) scores, indicating superior structural fidelity. Building on this, we further evaluate the performance of these four LLMs in generating sequences related to three other rare diseases, Landau Kleffner Syndrome (LKS) (Figure 6), Progressive Multifocal Leukoencephalopathy (PML) (Figure 7), and Paraneoplastic Neurologic Syndromes (PNS) (Figure 8), using the same strategy. Two summary tables showing the landscape of both global RMSD and LDDT between ground truth and predicted DNA structures can be found in Table 1 and Table 2.

**Lightweight and Efficient Adaptation.** To alleviate resource constraints, methods like quantization, model pruning (Cheng et al., 2024), knowledge refinement (Subagdja et al., 2024), and low-rank adaptation (LoRA) (Hu et al., 2021; Wang et al., 2024b,a) reduce model size while preserving performance. These scaffolds enable smaller labs to utilize prompt-based models more efficiently and accelerate model refinement.

## 6 Conclusion

In this survey, we examined how prompt-based methods enhance LLMs for biological sequence analysis, including applications in DNA, RNA, proteins, and drug discovery. Prompt engineering enables generalization in low-resource settings through zero- and few-shot learning. We outline three key directions for future research: data-centric prompting, unified multimodal integration, and scalable, efficient prompting. As LLMs evolve, these approaches will be pivotal in advancing precision medicine and computational biology, unlocking new opportunities for AI-driven bioinformatics.

## Limitations

This study is the first survey of recent advancements in biological sequence with language model prompting. We have made our best effort, but some limitations remain. We present recent methods and application domains rather than an exhaustive coverage. Due to space constraints, we can only provide brief method summaries without exhaustive technical details. Due to focusing primarily on publication from bioinformatics-related journals or conferences, we may have overlooked significant work published in other venues. We will continue to monitor the research community, incorporate new perspectives, and address any omissions in future updates.

In addition, we only use AI tools to polish the language of our paper.

## Ethics Statement

This paper does not involve ethics-related issues.

## References

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

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*.

Nicholas Ampazis and Flora Sakketou. 2024. Diversifying multi-head attention in the transformer model. *Machine Learning and Knowledge Extraction*, 6(4):2618–2638.

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

Duanhua Cao, Mingan Chen, Runze Zhang, Zhaokun Wang, Manlin Huang, Jie Yu, Xinyu Jiang, Zhehuan Fan, Wei Zhang, Hao Zhou, et al. 2024. Surfdock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nature Methods*, pages 1–13.

Yung-Chun Chang, Ming-Siang Huang, Yi-Hsuan Huang, and Yi-Hsuan Lin. 2025. The influence of prompt engineering on large language models for protein–protein interaction identification in biomedical literature. *Scientific Reports*, 15(1):15493.

Valerie Chen, Muyu Yang, Wenbo Cui, Joon Sik Kim, Ameet Talwalkar, and Jian Ma. 2024. Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. *Nature methods*, 21(8):1454–1461.

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. 2024. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carl Edwards, Aakanksha Naik, Tushar Khot, Martin Burke, Heng Ji, and Tom Hope. 2023. Synergpt: In-context learning for personalized drug synergy prediction and drug design. *arXiv preprint arXiv:2307.11694*.

Ziqi Gao, Xiangguo Sun, Zijing Liu, Yu Li, Hong Cheng, and Jia Li. 2024. Protein multimer structure prediction via prompt learning. *Preprint*, arXiv:2402.18813.

Louie Giray. 2023. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633.

Adam SL Graefe, Miriam R Hübner, Filip Rehburg, Steffen Sander, Sophie AI Klopfenstein, Samer Alkarkoukly, Ana Grönke, Annic Weyersberg, Daniel Danis, Jana Zschüntzsch, et al. 2025. An ontology-based rare disease common data model harmonising international registries, fhir, and phenopackets. *Scientific Data*, 12(1):234.

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xuanming Hu, Dongjie Wang, Wangyang Ying, and Yanjie Fu. 2024. Reinforcement feature transformation for polymer property performance prediction. In

*Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4538–4545.

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

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. *Preprint*, arXiv:2405.06649.

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

Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. 2024. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, pages 1–15.

Deqian Kong, Yuhao Huang, Jianwen Xie, Edouardo Honig, Ming Xu, Shuanghong Xue, Pei Lin, Sanping Zhou, Sheng Zhong, Nanning Zheng, et al. 2024. Dual-space optimization: Improved molecule sequence design by latent prompt transformer. *arXiv preprint arXiv:2402.17179*.

Zhongshen Li, Junru Jin, Wentao Long, and Leyi Wei. 2023. Plpmpro: Enhancing promoter sequence prediction with prompt-learning based pre-trained language model. *Computers in Biology and Medicine*, 164:107260.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, 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*.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. 2023. 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. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Fangliangzi Meng, Hongrun Zhang, Ruodan Yan, Guohui Chuai, Chao Li, and Qi Liu. 2024. Genomics-guided representation learning for pathologic pan-cancer tumor microenvironment subtype prediction. *Preprint*, arXiv:2406.06517.

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

Qizhi Pei, Lijun Wu, Jinhua Zhu, Yingce Xia, Shufang Xie, Tao Qin, Haiguang Liu, Tie-Yan Liu, and Rui Yan. 2023. Breaking the barriers of data scarcity in drug–target affinity prediction. *Briefings in Bioinformatics*, 24(6):bbad386.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. 2019. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*.

David B Searls. 1997. Linguistic approaches to biological sequences. *Bioinformatics*, 13(4):333–344.

Budhitama Subagdja, D Shanthoshigaa, Zhaoxia Wang, and Ah-Hwee Tan. 2024. Machine learning for refining knowledge graphs: A survey. *ACM Computing Surveys*, 56(6):1–38.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sheng Wang, Liheng Chen, Pengan Chen, Jingwei Dong, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2024a. Mos: Unleashing parameter efficiency of low-rank adaptation with mixture of shards. *arXiv preprint arXiv:2410.00938*.

Sheng Wang, Boyang Xue, Jiacheng Ye, Jiyue Jiang, Liheng Chen, Lingpeng Kong, and Chuan Wu. 2024b. Prolora: Partial rotation empowers more parameter-efficient lora. *arXiv preprint arXiv:2402.16902*.

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*.

Peng Wu, Huabin Du, Yingchao Yan, Tzong-Yi Lee, Chen Bai, and Song Wu. 2024. Guided diffusion for molecular generation with interaction prompt. *Briefings in Bioinformatics*, 25(3):bbae174.

Xi Xiao, Wentao Wang, Jiacheng Xie, Lijing Zhu, Gaofei Chen, Zhengji Li, Tianyang Wang, and Min Xu. 2024. Hgtdp-dta: Hybrid graph-transformer with dynamic prompt for drug-target binding affinity prediction. *arXiv preprint arXiv:2406.17697*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. 2024. Prompt engineering paradigms for medical applications: Scoping review. *Journal of Medical Internet Research*, 26:e60501.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42.

Ke Zhang, Yimiao Feng, and Jie Zheng. 2024a. Prompt-based generation of natural language explanations of synthetic lethality for cancer drug discovery. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13131–13142.

Qiang Zhang, Zeyuan Wang, Yuqiang Han, Haoran Yu, Xurui Jin, and Huajun Chen. 2022. Prompt-guided injection of conformation to pre-trained protein model. *Preprint*, arXiv:2202.02944.

Xinhao Zhang, Jinghan Zhang, Banafsheh Rekabdar, Yuanchun Zhou, Pengfei Wang, and Kunpeng Liu. 2024b. Dynamic and adaptive feature generation with llm. *arXiv preprint arXiv:2406.03505*.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. 2024c. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121.

Xiaosa Zhao, Kunpeng Liu, Wei Fan, Lu Jiang, Xiaowei Zhao, Minghao Yin, and Yanjie Fu. 2020. Simplifying reinforced feature selection via restructured choice strategy of single agent. In *2020 IEEE International conference on data mining (ICDM)*, pages 871–880. IEEE.

Wenxin Zhou and Thuy Hang Ngo. 2024. Using pretrained large language model with prompt engineering to answer biomedical questions. *arXiv preprint arXiv:2407.06779*.

Shuxian Zou, Shentong Mo, Hui Li, Xingyi Cheng, Le Song, and Eric Xing. 2023. Linker-tuning: Optimizing continuous prompts for heterodimeric protein prediction.

# A Appendix

## A.1 Literature Tree

## A.2 Additional Case Studies

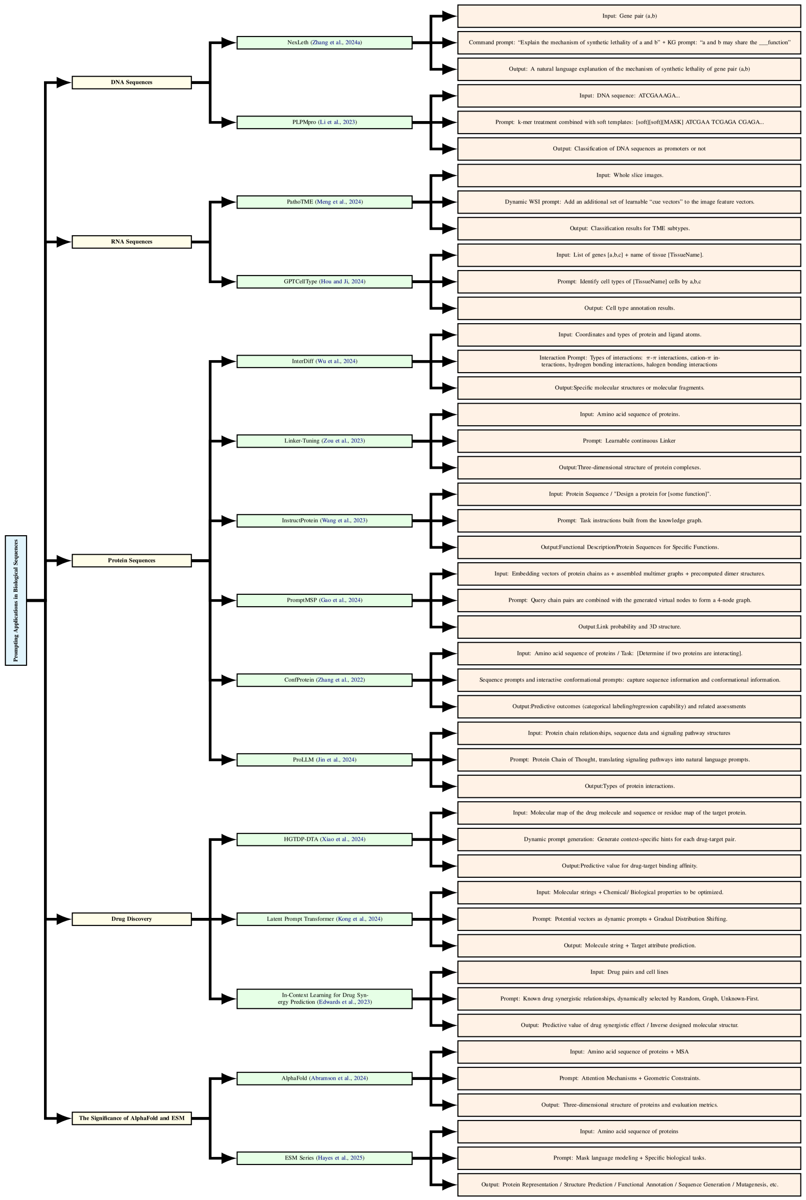

Prompting Applications in Biological Sequences
DNA Sequences
NexLeth (Zhang et al., 2024a)
Input: Gene pair (a,b)
Command prompt: “Explain the mechanism of synthetic lethality of a and b” + KG prompt: “a and b may share the ___function”
Output: A natural language explanation of the mechanism of synthetic lethality of gene pair (a,b)
PLPMpro (Li et al., 2023)
Input: DNA sequence: ATCGAAAGA...
Prompt: k-mer treatment combined with soft templates: [soft][soft][MASK] ATCGAA TCGAGA CGAGA...
Output: Classification of DNA sequences as promoters or not
RNA Sequences
PathoTME (Meng et al., 2024)
Input: Whole slice images.
Dynamic WSI prompt: Add an additional set of learnable “cue vectors” to the image feature vectors.
Output: Classification results for TME subtypes.
GPTCellType (Hou and Ji, 2024)
Input: List of genes [a,b,c] + name of tissue [TissueName].
Prompt: Identify cell types of [TissueName] cells by a,b,c
Output: Cell type annotation results.
Protein Sequences
InterDiff (Wu et al., 2024)
Input: Coordinates and types of protein and ligand atoms.
Interaction Prompt: Types of interactions: $\pi$-$\pi$ interactions, cation-$\pi$ interactions, hydrogen bonding interactions, halogen bonding interactions
Output:Specific molecular structures or molecular fragments.
Linker-Tuning (Zou et al., 2023)
Input: Amino acid sequence of proteins.
Prompt: Learnable continuous Linker
Output:Three-dimensional structure of protein complexes.
InstructProtein (Wang et al., 2023)
Input: Protein Sequence / "Design a protein for [some function]".
Prompt: Task instructions built from the knowledge graph.
Output:Functional Description/Protein Sequences for Specific Functions.
PromptMSP (Gao et al., 2024)
Input: Embedding vectors of protein chains as + assembled multimer graphs + precomputed dimer structures.
Prompt: Query chain pairs are combined with the generated virtual nodes to form a 4-node graph.
Output:Link probability and 3D structure.
ConfProtein (Zhang et al., 2022)
Input: Amino acid sequence of proteins / Task: [Determine if two proteins are interacting].
Sequence prompts and interactive conformational prompts: capture sequence information and conformational information.
Output:Predictive outcomes (categorical labeling/regression capability) and related assessments
ProLLM (Jin et al., 2024)
Input: Protein chain relationships, sequence data and signaling pathway structures
Prompt: Protein Chain of Thought, translating signaling pathways into natural language prompts.
Output:Types of protein interactions.
Drug Discovery
HGTDP-DTA (Xiao et al., 2024)
Input: Molecular map of the drug molecule and sequence or residue map of the target protein.
Dynamic prompt generation: Generate context-specific hints for each drug-target pair.
Output:Predictive value for drug-target binding affinity.
Latent Prompt Transformer (Kong et al., 2024)
Input: Molecular strings + Chemical/ Biological properties to be optimized.
Prompt: Potential vectors as dynamic prompts + Gradual Distribution Shifting.
Output: Molecule string + Target attribute prediction.
In-Context Learning for Drug Synergy Prediction (Edwards et al., 2023)
Input: Drug pairs and cell lines
Prompt: Known drug synergistic relationships, dynamically selected by Random, Graph, Unknown-First.
Output: Predictive value of drug synergistic effect / Inverse designed molecular structur.
The Significance of AlphaFold and ESM
AlphaFold (Abramson et al., 2024)
Input: Amino acid sequence of proteins + MSA
Prompt: Attention Mechanisms + Geometric Constraints.
Output: Three-dimensional structure of proteins and evaluation metrics.
ESM Series (Hayes et al., 2025)
Input: Amino acid sequence of proteins
Prompt: Mask language modeling + Specific biological tasks.
Output: Protein Representation / Structure Prediction / Functional Annotation / Sequence Generation / Mutagenesis, etc.

Figure 4: Hierarchical taxonomy of prompt-based applications in biological sequence modeling.

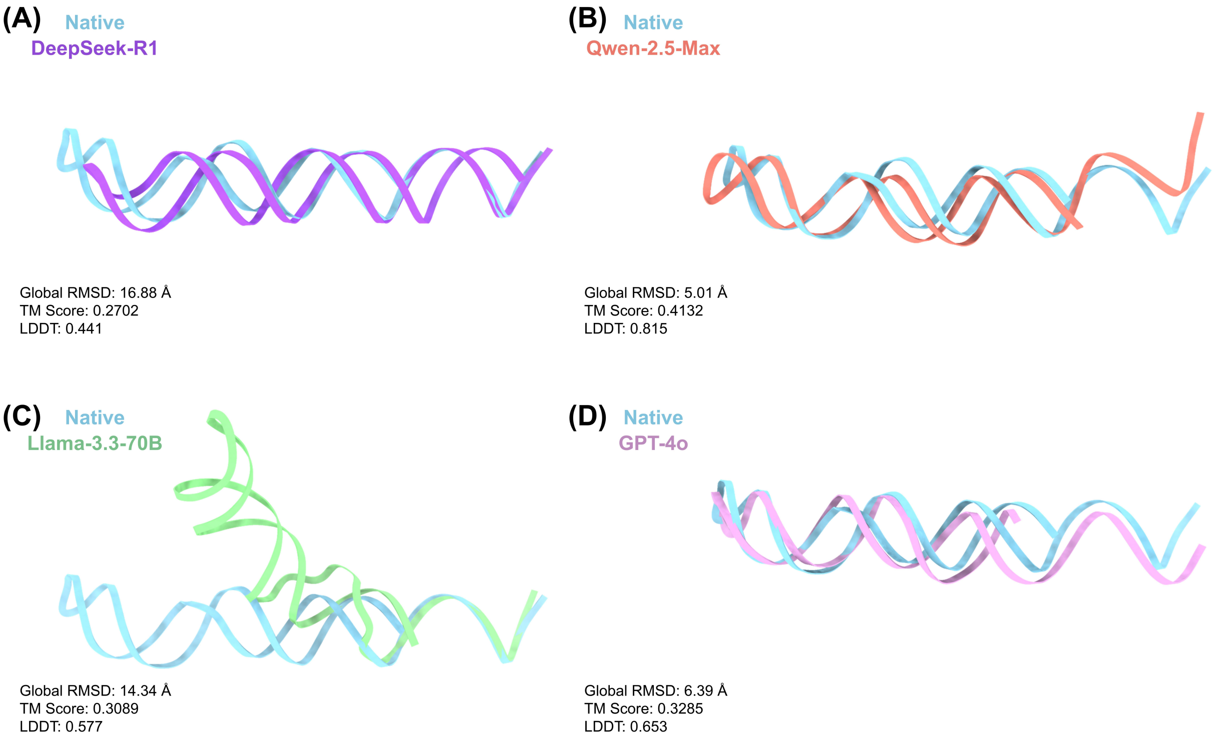


Figure 5: Structural alignment between the ground truth structure (light blue), a 70 bp DNA related to Tardive Dyskinesia (TD) biogenesis, and the predicted structures based on prompt-based generated DNA sequences: (**a**) DeepSeek-R1 (purple), Global RMSD = 16.88 Å, TM Score = 0.2702, LDDT = 0.441. (**b**) Qwen-2.5-Max (salmon), Global RMSD = 5.01 Å, TM Score = 0.4132, LDDT = 0.815. (**c**) Llama-3.3-70B (light green), Global RMSD = 14.34 Å, TM Score = 0.3089, LDDT = 0.577. (**d**) GPT-4o (lavender), Global RMSD = 6.39 Å, TM Score = 0.3285, LDDT = 0.653. (Reference sequence: GATAGAGAGACAAAGAGGAAAAGAGAGCGAGGTAGAAAACGGAT-ACTGCCTATGCCTACTCCATCCCTCT)

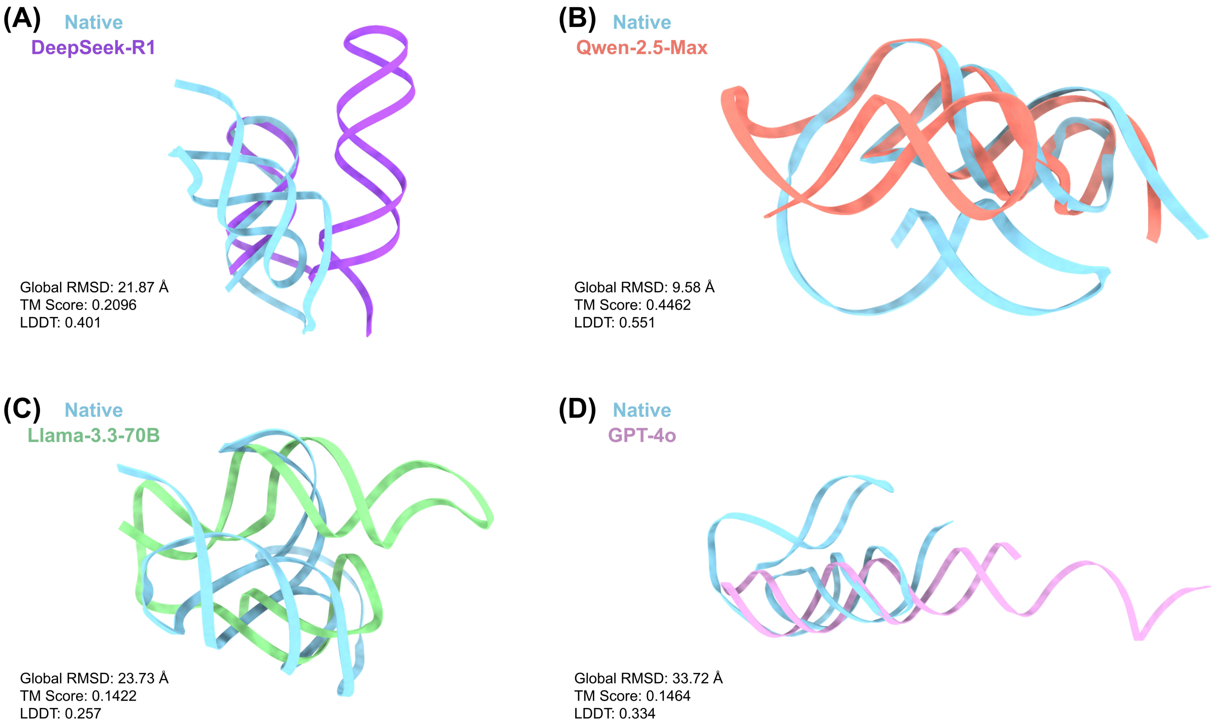


Figure 6: Structural alignment between the ground truth structure (light blue), a 70 bp DNA related to Landau Kleffner Syndrome (LKS) biogenesis, and the predicted structures based on prompt-based generated DNA sequences: (**a**) DeepSeek-R1 (purple), Global RMSD = 21.87 Å, TM Score = 0.2096, LDDT = 0.401. (**b**) Qwen-2.5-Max (salmon), Global RMSD = 9.58 Å, TM Score = 0.4462, LDDT = 0.551. (**c**) Llama-3.3-70B (light green), Global RMSD = 23.73 Å, TM Score = 0.1422, LDDT = 0.257. (**d**) GPT-4o (lavender), Global RMSD = 33.72 Å, TM Score = 0.1464, LDDT = 0.334. (Reference sequence: CTCTTTCTCTCCCTACCTCCCTCGCTCAGCAGCTCCCG-GTCGCACAACTCCCAGCAGCCGGCGCTGGGGA)

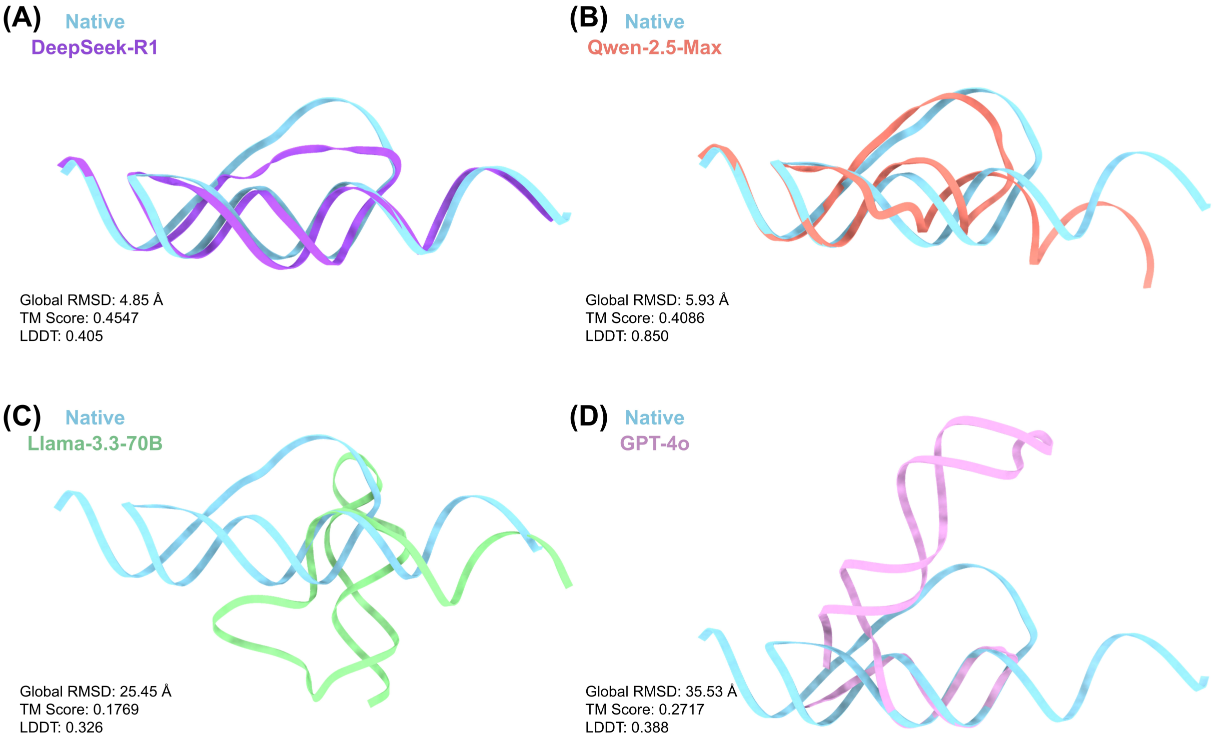


Figure 7: Structural alignment between the ground truth structure (light blue), a 70 bp DNA related to Progressive Multifocal Leukoencephalopathy (PML) biogenesis, and the predicted structures based on prompt-based generated DNA sequences: (**a**) DeepSeek-R1 (purple), Global RMSD = 4.85 Å, TM Score = 0.4547, LDDT = 0.405. (**b**) Qwen-2.5-Max (salmon), Global RMSD = 5.93 Å, TM Score = 0.4086, LDDT = 0.850. (**c**) Llama-3.3-70B (light green), Global RMSD = 25.45 Å, TM Score = 0.1769, LDDT = 0.326. (**d**) GPT-4o (lavender), Global RMSD = 35.53 Å, TM Score = 0.2717, LDDT = 0.388. (Reference sequence: CCAAAGGCTAGATTTAAAAACCCCAAAT-GTGCAATCTGGTGAATTTATAGAAAGAAGTATTGCACCAGGA)

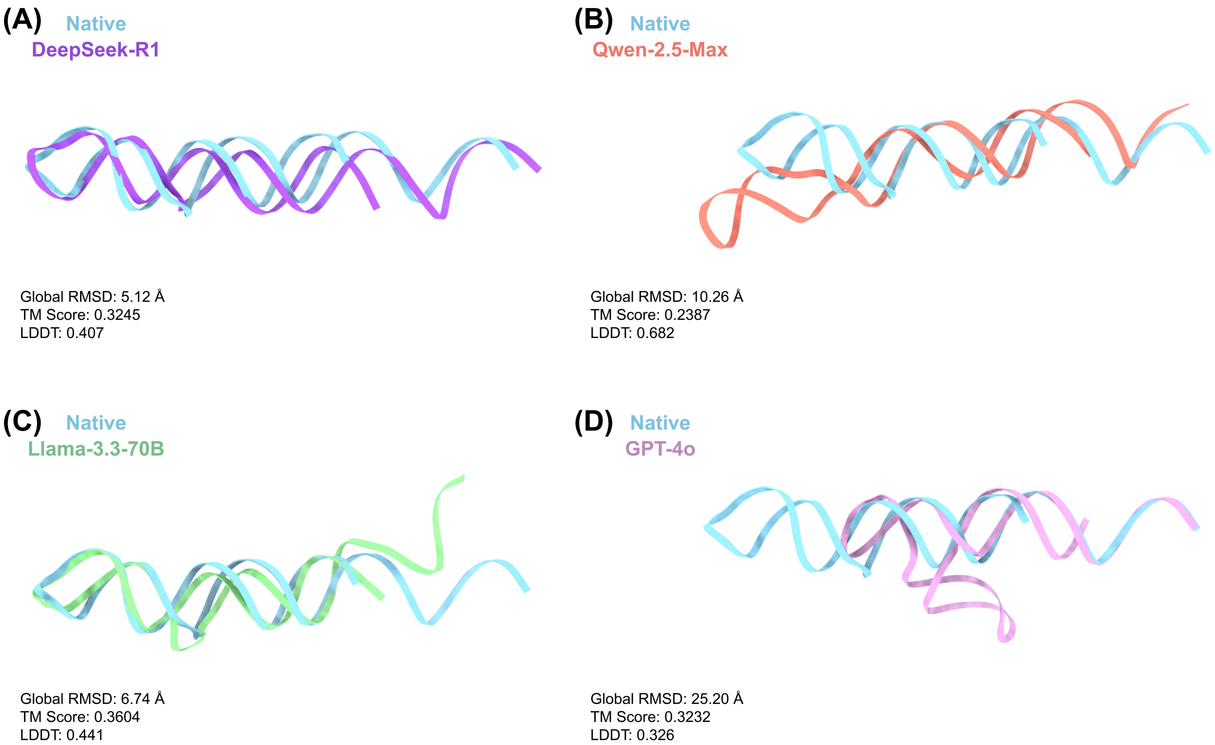


Figure 8: Structural alignment between the ground truth structure (light blue), a 70 bp DNA related to Paraneoplastic Neurologic Syndromes (PNS) biogenesis, and the predicted structures based on prompt-based generated DNA sequences: (**a**) DeepSeek-R1 (purple), Global RMSD = 5.12 Å, TM Score = 0.3245, LDDT = 0.407. (**b**) Qwen-2.5-Max (salmon), Global RMSD = 10.26 Å, TM Score = 0.2387, LDDT = 0.682. (**c**) Llama-3.3-70B (light green), Global RMSD = 6.74 Å, TM Score = 0.3604, LDDT = 0.441. (**d**) GPT-4o (lavender), Global RMSD = 25.20 Å, TM Score = 0.3232, LDDT = 0.326. (Reference sequence: AGCAGACGCTCCCTCAGCAAGGACAGCAGAG-GACCAGCTAAGAGGGAGAGAAGCAACTACAGACCCCCCC)

| **LLM** \ **Disease** | LKS | PML | PNS | TD |
|---|---|---|---|---|
| DeepSeek-R1 | 22.77 | 24.21 | 11.14 | 9.94 |
| Qwen-2.5-Max | **20.68** | 35.45 | 15.98 | **4.82** |
| Llama-3.3-70B | 27.79 | **22.57** | **9.72** | 8.68 |
| GPT-4o | 25.03 | 26.29 | 13.67 | 11.48 |

Table 1: Landscape of global **RMSD** (**Å**) between ground truth and predicted DNA structures for a 70 bp sequence associated with the biogenesis of the four rare diseases listed above. The predicted structures were generated by four LLMs prompted with the same biological query. For each LLM and disease, **n = 10** sequences were generated, and each cell reports the **mean RMSD** across these predictions. The value of RMSD is always used to quantify the structural divergence between each predicted DNA structure and the ground truth, with **lower values** indicating **higher structural fidelity**. **Bolded values** denote the lowest RMSD in each disease column, highlighting the most accurate prediction per condition.

| **LLM** \ **Disease** | LKS | PML | PNS | TD |
|---|---|---|---|---|
| DeepSeek-R1 | **0.457** | 0.519 | 0.523 | 0.595 |
| Qwen-2.5-Max | 0.388 | 0.318 | **0.592** | **0.863** |
| Llama-3.3-70B | 0.361 | 0.327 | 0.517 | 0.540 |
| GPT-4o | 0.451 | **0.556** | 0.543 | 0.527 |

Table 2: Landscape of **LDDT** between ground truth and predicted DNA structures for a 70 bp sequence associated with the biogenesis of the four rare diseases listed above. The predicted structures were generated by four LLMs prompted with the same biological query. For each LLM and disease, **n = 10** sequences were generated, and each cell reports the **mean LDDT** across these predictions. The value of LDDT, ranging from 0 to 1, is always used to quantify the structural divergence between each predicted DNA structure and the ground truth, with **higher values** indicating **higher structural fidelity**. **Bolded values** denote the highest LDDT in each disease column, highlighting the most accurate prediction per condition.