

# Lost in Tokenization: Context as the Key to Unlocking Biomolecular Understanding in Scientific LLMs

Kai Zhuang<sup>+1,2,3</sup>, Jiawei Zhang<sup>+2</sup>, Yumou Liu<sup>+4</sup>, Hanqun Cao<sup>5</sup>, Chunbin Gu<sup>5</sup>, Mengdi Liu<sup>6</sup>, Zhangyang Gao<sup>1</sup>, Zitong Jerry Wang<sup>2</sup>, Xuanhe Zhou<sup>4</sup>, Pheng-Ann Heng<sup>5</sup>, Lijun Wu<sup>1</sup>, Conghui He<sup>1</sup>, Cheng Tan<sup>1,5</sup>

<sup>1</sup>Shanghai Artificial Intelligence Laboratory, <sup>2</sup>Westlake University, <sup>3</sup>Shanghai Innovation Institute, <sup>4</sup>Shanghai Jiaotong University, <sup>5</sup>The Chinese University of Hong Kong, <sup>6</sup>Institute of Computing Technology, Chinese Academy of Sciences

Scientific Large Language Models (Sci-LLMs) have emerged as a promising frontier for accelerating biological discovery. However, these models face a fundamental challenge when processing raw biomolecular sequences: the *tokenization dilemma*. Whether treating sequences as a specialized language, risking the loss of functional motif information, or as a separate modality, introducing formidable alignment challenges, current strategies fundamentally limit their reasoning capacity. We challenge this sequence-centric paradigm by positing that a more effective strategy is to provide Sci-LLMs with high-level structured context derived from established bioinformatics tools, thereby bypassing the need to interpret low-level noisy sequence data directly. Through a systematic comparison of leading Sci-LLMs on biological reasoning tasks, we tested three input modes: sequence-only, context-only, and a combination of both. Our findings are striking: the context-only approach consistently and substantially outperforms all other modes. Even more revealing, the inclusion of the raw sequence alongside its high-level context consistently degrades performance, indicating that raw sequences act as informational noise, even for models with specialized tokenization schemes. These results suggest that the primary strength of existing Sci-LLMs lies not in their nascent ability to interpret biomolecular syntax from scratch, but in their profound capacity for reasoning over structured, human-readable knowledge. Therefore, we argue for reframing Sci-LLMs not as sequence decoders, but as powerful reasoning engines over expert knowledge. This work lays the foundation for a new class of hybrid scientific AI agents, repositioning the developmental focus from direct sequence interpretation towards high-level knowledge synthesis.

**Date:** October 31, 2025

**Correspondence:** Conghui He, [heconghui@pjlab.org.cn](mailto:heconghui@pjlab.org.cn), Cheng Tan, [tancheng@pjlab.org.cn](mailto:tancheng@pjlab.org.cn)

 [Code](#)  [Website](#)  [Dataset](#)

## 1 Introduction

The convergence of artificial intelligence and the life sciences has given rise to a new class of powerful tools: Scientific Large Language Models (Sci-LLMs). Built on Transformer architectures (e.g. BERT, GPT) that have revolutionized natural language processing [14], these models hold immense promise for accelerating biological discovery [22]. From predicting protein function [8] to designing novel therapeutics [16], Sci-LLMs such as Intern-S1 [5], Evolla [38], and NatureLM [34] are being developed to interpret the complex “language of life” encoded in DNA, RNA, and protein sequences [32]. Early efforts have demonstrated their potential, sparking visions of an AI-driven future for scientific research. This burgeoning field has largely coalesced around two primary strategies for integrating biomolecular data [19]. The first “*sequence-as-language*” approach treats sequences as a specialized form of language,

extending the model’s vocabulary to include individual amino acids or nucleotides and pre-training it on vast corpora of sequence and text data. The second “*sequence-as-modality*” approach, inspired by multimodal learning, treats sequences as a distinct modality, employing a specialized encoder (e.g., a pre-trained biological foundation model like ESM [23] and Evo [12]) to generate rich embeddings that are then aligned with and injected into the language model’s input space, allowing LLMs to reason over high-level features of the sequence provided by the encoder, rather than the raw sequence itself [1, 25, 10].

While both paradigms have shown progress, they share a fundamental, yet often overlooked, vulnerability that we term the *tokenization dilemma*. In the “*sequence-as-language*” paradigm, the tokenization process is often too granular [29, 8]. By breaking down sequences into their atomic components—single amino acids or nucleotides—it destroys the very structures that carry biological meaning: functional motifs, domains, and regulatory elements [14]. The model is consequently forced into the complicated task of re-learning these fundamental “words” of biology from a stream of disconnected “letters,” a process that is both inefficient and struggles with generalization. Conversely, the “*sequence-as-modality*” paradigm, while preserving structural information within its high-fidelity embeddings, introduces a formidable alignment challenge [17]. The hidden space learned by a bioinformatics encoder is governed by the principles of evolution and biophysics, a world of alpha-helices and selective pressure. The hidden space of an LLM, however, is shaped by human language. Bridging this profound semantic gap between the two modalities is a non-trivial task, and imperfect alignment can introduce ambiguity or even misinterpretation, limiting the model’s ability to ground its reasoning accurately in the underlying biological reality. We are, in essence, asking these models to perform a task for which they are ill-equipped: they are becoming lost in tokenization.

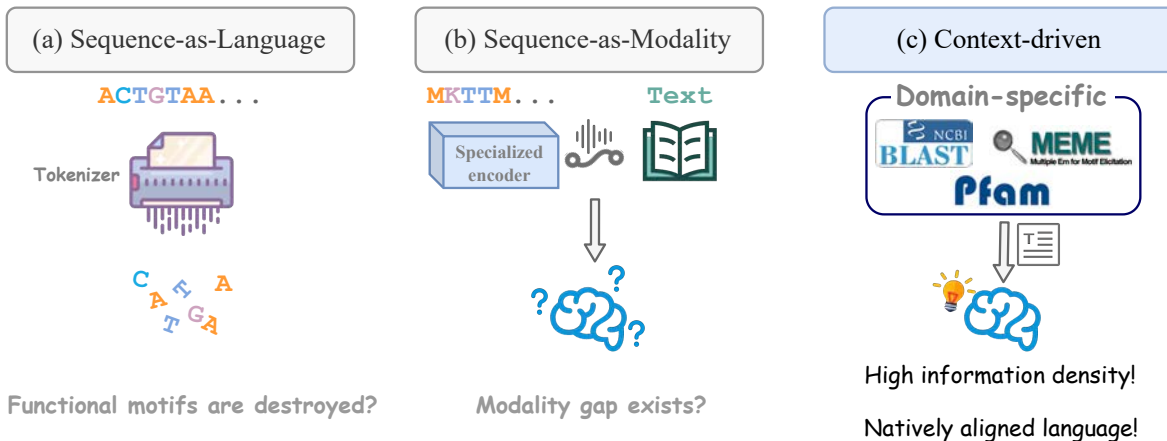


Figure 1: Paradigms for integrating biomolecular sequences into Sci-LLMs. (a) The sequence-as-language approach, tokenization fragments sequences into atomic symbols, potentially destroying functional motifs. (b) The sequence-as-modality approach preserves structure via specialized encoders but suffers from semantic misalignment with natural language. (c) The context-driven approach leverages bioinformatics tools to provide high-density, natively aligned textual context.

In this work, we challenge the prevailing sequence-centric view and propose an alternative, more effective paradigm to overcome the tokenization dilemma. We hypothesize that rather than forcing LLMs to directly decipher the noisy, low-level syntax of raw biomolecular sequences, we should leverage their core strength: reasoning over high-level, structured knowledge. Decades of accumulated biological wisdom are embedded in expert tools and databases – from BLAST for sequence homology to Pfam for conserved domains and Gene Ontology for functional terms. As shown in Figure 1, we posit that these resources can be transformed into an information-rich textual context for the LLM. This “context”, presented as human-readable text, is not only information-dense, having already distilled

functional insights from the raw sequence, but is also natively aligned with the LLM’s linguistic domain, entirely circumventing the tokenization dilemma.

We conduct a systematic empirical study across a representative set of state-of-the-art Sci-LLMs. Surprisingly, we observe that adding the raw sequence to an already informative context often degrades performance: the sequence acts as a form of “informational noise” that confuses an otherwise well-informed model. When both sequence and context are given, the sequence introduces misleading signals that reduce accuracy, suggesting that the true power of current Sci-LLMs lies not in their ability to serve as *de novo* sequence interpreters, but as sophisticated reasoning engines over integrated domain knowledge. Models that are fed high-level biological context can make insightful connections and generalizations whereas those fed only raw sequences struggle to draw any inference until they essentially “learn biology” from scratch.

## 2 Related Work

### 2.1 Foundation Models in Biological Representation

Foundation models for biological sequences have made rapid strides in representation learning. In the protein domain, large language models like ProtBERT [11] and the ESM series [23, 15] are trained on massive sequence corpora, capturing signals of evolutionary conservation, structural motifs, and residue co-variation that enable downstream generalization. On the nucleotide side, models such as DNABERT [20] and the more recent Nucleotide Transformer [9] apply *k*-mer tokenization or other subword strategies to genome-scale data, achieving high accuracy in identifying promoters, splice sites, and transcription factor binding locations. Multi-species genome models like DNABERT-2 [39] further improve efficiency by replacing *k*-mers with Byte-Pair Encoding to accommodate longer input sequences. Meanwhile, specialized transformer architectures have extended context lengths to capture distal regulatory interactions and boost gene expression prediction [4, 27, 28]. Despite their powerful representational capacity, these bio-sequence foundation models largely act as “black boxes”. Their internal embeddings are high-dimensional and not straightforwardly mapped to human-interpretable biological units like motifs, domains, or pathways, making it difficult to extract mechanistic insight.

### 2.2 Scientific Large Language Models

Large language models tailored to scientific domains (Sci-LLMs) have rapidly advanced, extending the success of general LLMs into tasks like protein or molecule design, genomic analysis, and scientific reasoning. Galactica [32], a 120-billion-parameter model trained on a corpus of papers and knowledge bases, was introduced to store and reason over scientific knowledge. Domain-focused sequence models have also emerged: NatureLM [34], for example, is a unified sequence-based model pre-trained across proteins, nucleic acids and small molecules. Likewise, Intern-S1 [5] is a recent large multimodal MoE model (28B activated parameters) with specialized tokenization and encoders for different scientific modalities. In this work, we focus on biomolecular understanding as a representative scientific challenge: information is inherently encoded in sequences (genes or proteins), which can be expressed in textual form or as a distinct modality, making it an ideal testbed for probing how well Sci-LLMs integrate domain knowledge and whether they truly understand biological sequences.

### 2.3 Existing Strategies in Bridging Sequences and Language

Sci-LLMs have adopted several strategies to bridge low-level biomolecular sequences with higher-level reasoning and knowledge. One common approach is treating sequences as a specialized language. Models like NatureLM [34] and Intern-S1 [5] ingest raw or tokenized sequences directly as input, training on vast datasets of sequences annotated with text so that the model learns joint representations. Another emerging strategy is treating sequences as a separate modality. For example, EvoLLaMA [25]

incorporates a protein structure encoder and a sequence encoder alongside an LLM to enable multimodal protein question-answering, and Evolla [38] employs SaProt [30] as the structure encoder. BioReason [12] similarly couples a frozen DNA foundation model Evo [28] with a language model Qwen3 [35], so that genomic sequences are converted into contextual embeddings which the LLM can reason over in natural language. A third line of work explores agent-based or tool-augmented approaches. Rather than having a single model directly analyze sequences, the LLM is equipped with the ability to call external tools or databases as needed. Notable examples include GeneAgent [33], which self-verifies for gene-set analysis using domain databases, and ChemCrow [7], which uses an agent to plan multi-step chemistry tasks by invoking a suite of expert tools. While all these strategies have pushed the frontier of scientific AI [18], it remains unclear how much of the success in Sci-LLMs comes from genuine reasoning over raw sequences. In this work, we adopt a deliberately context-driven baseline—providing the model with only high-level, structured annotations of the sequence. By comparing this setup to one where the model sees the raw sequence, we can assess how and when sequence information truly adds value.

### 3 Preliminaries

#### 3.1 The Biomolecular Understanding Task

Let  $\mathcal{S}$  be the space of all possible biomolecular sequences (e.g., protein, RNA, DNA, and small molecules),  $\mathcal{Q}$  be the space of natural language questions about a sequence, and  $\mathcal{A}$  be the space of plausible natural language answers. The general task is to learn a function  $f : \mathcal{S} \times \mathcal{Q} \rightarrow \mathcal{A}$  that maps a sequence  $s \in \mathcal{S}$  and a question  $q \in \mathcal{Q}$  to a factually correct and relevant answer  $a \in \mathcal{A}$ .

A Scientific LLM, denoted as  $\mathcal{M}$ , aims to approximate this function by learning a set of optimal parameters  $\theta$ . The generation of an answer can be expressed as:

$$a = \mathcal{M}(s, q; \theta) \tag{1}$$

The fundamental distinction between the paradigms we investigate lies in how the sequence  $s$  and question  $q$  are represented and processed by the model  $\mathcal{M}$ .

#### 3.2 Sequence-as-Language

This approach, utilized by models such as NatureLM [34] and Intern-S1 [5], treats a biomolecular sequence as a specialized string of text. Let  $T_{seq}$  be a tokenizer that maps a sequence  $s$  into a series of tokens from a biological vocabulary,  $V_{bio}$ , and let  $T_{text}$  be a standard tokenizer for a natural language question  $q$  with vocabulary  $V_{text}$ . The model operates on an extended vocabulary  $V_{ext} = V_{text} \cup V_{bio}$ . The input to the LLM,  $X_{input}$ , is formed by the concatenation of the tokenized question and sequence:

$$X_{input} = [T_{text}(q); T_{seq}(s)] \tag{2}$$

The model  $\mathcal{M}$  then processes this unified token sequence autoregressively to generate the answer  $a$ :

$$P(a|s, q) = \prod_{k=1}^{|a|} P(a_k | a_{<k}, X_{input}; \theta) \tag{3}$$

It introduces the first horn of the tokenization dilemma: the **weak representation** comes from the low-level tokenization atomizes the sequence, destroying the hierarchical structures of functional motifs. The model receives a high-dimensional but low-information-density signal, from which it must re-learn the fundamental grammar of biology, a notoriously difficult and data-intensive task.

### 3.3 Sequence-as-Modality

Inspired by successes in vision-language modeling, this paradigm—employed by models like Evolla [38] and BioReason [12]—treats the biomolecular sequence as a distinct, non-textual modality. A specialized, pre-trained biomolecular encoder,  $\mathcal{E}_{bio} : \mathcal{S} \rightarrow \mathbb{R}^{L \times d}$ , first transforms the sequence  $s$  into a sequence of rich, contextualized embeddings. An alignment module,  $\mathcal{A}_{align}$ , then projects these biological embeddings into the LLM’s semantic space, creating an aligned sequence representation  $E_{aligned.seq} \in \mathbb{R}^{K \times d}$ . The final input to the LLM is a structured combination of the embedded text and the aligned sequence embeddings:

$$X_{input} = [T_{text}(q); E_{aligned.seq}] \quad (4)$$

While this approach preserves the sequence’s structural integrity, it introduces the second horn of the tokenization dilemma: the challenge of **semantic misalignment**. The semantic space of  $\mathcal{E}_{bio}$  is governed by the principles of biophysics and evolution, whereas the LLM’s space is structured by human linguistics and logic. The alignment module  $\mathcal{A}_{align}$  must learn to bridge this profound semantic gap. Any imperfection in this translation can inject ambiguity or noise.

## 4 The Context-Driven Approach

In this work, we propose and investigate a third paradigm that circumvents the tokenization dilemma entirely. This approach posits that the most effective way to leverage an LLM is to provide it with what it processes best: high-quality, human-readable text.

We define a set of established bioinformatics tools as a function  $\mathcal{C} : \mathcal{S} \rightarrow \mathcal{T}_{context}$ , where  $\mathcal{T}_{context}$  is the space of structured, human-readable textual descriptions. This function transforms a raw sequence  $s$  into a high-level context  $c = \mathcal{C}(s)$ . The model’s input deliberately omits the raw sequence  $s$ :

$$X_{input} = [T_{text}(q); T_{text}(c)] \quad (5)$$

The model approximates the answer’s probability by conditioning only on high-level knowledge:

$$P(a|s, q) \approx P(a|c, q) = \prod_{k=1}^{|a|} P(a_k | a_{<k}, X_{input}; \theta) \quad (6)$$

This paradigm reframes the task from one of low-level sequence interpretation to one of high-level knowledge synthesis. The context  $c$  is information-dense and natively aligned with the LLM’s natural language space, shifting the model’s role from low-level sequence interpretation to high-level knowledge synthesis and reasoning.

Specifically, we design a pipeline to generate and structure the context for any given protein sequence. First, we generate a comprehensive functional profile by executing a multi-source toolchain. InterProScan [21] is used to identify conserved domains and motifs based on the sequence’s intrinsic features, while BLASTp [2] retrieves annotations from close homologs in the Swiss-Prot database [6]. For novel orphan sequences lacking hits from these tools, we use the tri-modal retrieval model ProTrek [31] as a fallback to generate a basic semantic description. The outputs from these tools are then integrated into a final context using an empirically-driven hierarchical strategy. The details are in the Appendix A.

#### Structured Prompt for Context-Driven Reasoning

You are a senior systems biologist. Analyze the input information to answer the given question.

-----

```

Question:[User’s Question Text]
-----
Conserved Domains (from Pfam):
[FOR EACH Pfam entry IN Pfam]:
- {the description of detected conserved domains/motifs}
Functional Annotations (from Homology via BLASTp):
- GO terms associated with the homolog:
- {the GO terms of the homolog}
Fallback Semantic Analysis (from ProTrek):
[ONLY if no homology or domain data is available]
[FOR EACH ProTrek entry In Protrek]:
- {the description of Protrek}
-----
Answer:{answer}

```

A central concern in fair evaluation is the prevention of information leakage. Our context-driven approach is explicitly designed to avoid label leakage along two complementary axes:

**Intrinsic analysis rather than identity lookup.** We employ InterProScan to detect conserved domains and motifs intrinsic to the query sequence. This constitutes an *ab initio*, feature-based analysis grounded in domain knowledge bases, not in annotation records of the query protein itself. Consequently, even for genuinely novel proteins, recognizable elements such as a kinase domain can be identified without ground-truth labels.

**Homology-based inference rather than direct annotation matching.** When using BLASTp, we restrict our context-driven approach to reading GO annotations from the homologous sequences, rather than from the query protein’s own record. This reflects standard bioinformatics practice: predicting the function of unknown sequences by analogy to characterized homologs rather than simply retrieving pre-annotated answers.

## 5 The Tokenization Dilemma in Practice

### 5.1 The Primacy of Context over Sequence

Following a standardized protocol inspired by Evolla [38], our benchmark focuses on three fundamental aspects of protein biology: molecular function, metabolic pathway involvement, and subcellular localization. For each protein in our test set, we generated queries corresponding to these categories (e.g., “What is the function of this protein?”). To ensure a set of factually grounded and verifiable ground truths, a question was only included if its corresponding annotation field was explicitly present in the source database entry, from which the answer was directly excerpted. Performance was quantified using an automated pipeline, leveraging a general-purpose LLM as an expert judge, a metric we term the LLM-Score. A detailed description of the dataset construction, evaluation protocol, and prompt design is provided in Appendices B and C.

We evaluate the performance of both specialized Sci-LLMs and leading general-purpose LLMs across three distinct input configurations: (i) Sequence-Only, where the model receives only the raw protein sequence; (ii) Sequence + Context, a combined input; (iii) Context-Only, where the model receives only the high-level context. The results are presented in Table 1.

## Context as the Key to Unlocking Biomolecular Understanding in Scientific LLMs

Table 1: Comparison of performance across specialized Sci-LLMs and general-purpose LLMs on our protein QA benchmark. ✓ indicates that the corresponding input modality was provided to the model. Results are reported on three task-specific subsets—*Function* (Func.), *Pathway* (Path.), and *Subcellular Location* (Sub. Loc.)—as well as the overall average (All). The best score for each model is underlined, and the overall best performance across all models is highlighted in bold.

Model	Sequence	Context	Func.	Path.	Sub. Loc.	All
<b><i>Specialized Sci-LLMs</i></b>						
Intern-S1	✓		20.57	26.56	69.75	43.33
Intern-S1	✓	✓	74.18	98.85	93.00	84.03
Intern-S1		✓	76.22	97.60	95.60	<u>86.15</u>
Evolla	✓		40.23	72.71	79.76	59.93
Evolla	✓	✓	57.46	84.69	83.05	70.53
Evolla		✓	65.77	83.33	81.88	<u>74.02</u>
NatureLM	✓		3.58	5.52	10.45	6.82
NatureLM	✓	✓	42.33	64.25	32.30	38.86
NatureLM		✓	44.77	51.35	32.51	<u>39.50</u>
<b><i>General LLMs</i></b>						
Deepseek-v3	✓		10.98	24.54	74.72	40.77
Deepseek-v3	✓	✓	77.40	91.35	94.75	<u>86.03</u>
Deepseek-v3		✓	75.79	93.96	93.65	84.99
Gemini2.5 Pro	✓		10.40	13.85	77.58	41.25
Gemini2.5 Pro	✓	✓	79.12	94.17	94.65	86.98
Gemini2.5 Pro		✓	79.17	98.65	94.56	<b><u>87.19</u></b>
GPT-5	✓		19.64	17.08	64.15	39.83
GPT-5	✓	✓	79.89	89.48	71.30	<u>76.45</u>
GPT-5		✓	77.25	85.73	73.05	75.76

**Takeaway:** Raw biomolecular sequences, when provided alone, offer limited utility and, when combined with context, consistently act as informational noise.

Our findings demonstrate that the Context-Only approach is dramatically superior, confirming our hypothesis: *LLMs excel when they can leverage their core strength of reasoning over structured knowledge*. Even more revealing is the consistent performance degradation observed in the Sequence + Context configuration. The inclusion of the raw sequence alongside its high-level summary resulted in a lower score. For instance, Evolla’s score dropped from 74.02 to 70.53, and Intern-S1’s from 86.15 to 84.03. This counter-intuitive result provides evidence that raw sequences, in their current tokenized form, are not merely redundant but actively detrimental, acting as a source of noise. The models become, as we posited, “lost in tokenization”. This phenomenon underscores the profound limitations of existing sequence tokenization paradigms.

## 5.2 Deconstructing the Dilemma I: The Weak Representation

We visualize the embeddings of the outputs, where ground-truth classes were established by clustering homologous proteins using MMseqs2 at a 50% sequence identity threshold. For each model, we extracted the final-layer embeddings for their outputs. We employed t-SNE [26] to project them into a 2D space. The quality of the resulting functional separation was then quantified by performing clustering on the high-dimensional embeddings and calculating the Adjusted Rand Index (ARI) against the MMseqs2 ground-truth clusters. For our context-driven approach, we generated embeddings from the structured context itself using the text embedding model Qwen-embedding [37]. The results are visualized in Figure 2.

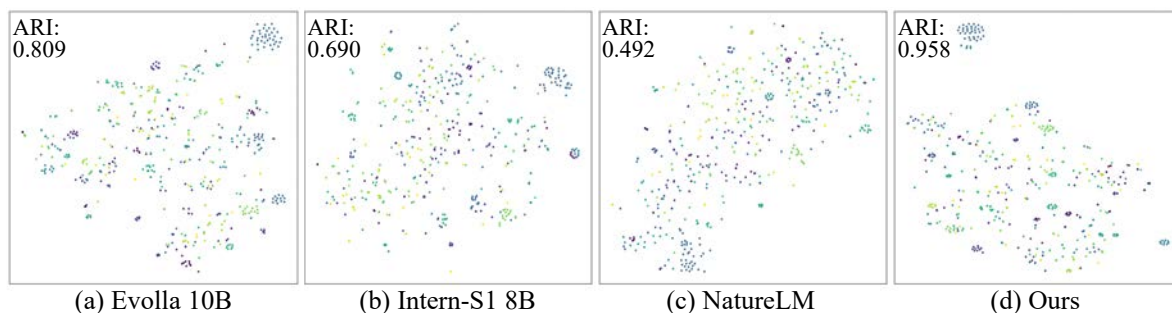


Figure 2: The visualization of representation spaces.

**Takeaway:** Simple context provides a vastly superior functional representation of proteins compared to both sequence-to-language/modality strategies.

The visualizations confirm the weak representation horn of the tokenization dilemma. The sequence-as-language models, NatureLM (c) and Intern-S1 (b), exhibit highly disorganized latent spaces, quantitatively confirmed by their low ARI scores of 0.492 and 0.690, respectively. Evolla (a), which employs the sequence-as-modality paradigm, demonstrates a significantly more structured representation, highlighting the benefit of using a specialized sequence encoder. However, both paradigms are dramatically outperformed by our context-driven approach (d). The representation derived purely from the textual context achieves near-perfect functional separation.

### 5.3 Deconstructing the Dilemma II: The Semantic Misalignment

While the sequence-as-modality paradigm, exemplified by Evolla, overcomes the weak representation problem, it introduces a more subtle yet equally critical challenge: semantic misalignment. The specialized encoder and the generalist LLM operate in fundamentally different semantic worlds—one governed by biophysics, the other by linguistics. We performed a layer-wise representational analysis of the Evolla-10B model, tracing the informational journey of a protein sequence from its biological embedding to its final interpretation by the language model. As shown in Figure 3, the initial SaProt encoder generates a well-structured latent space. As the Q-Former works to translate these biological embeddings for the LLM, the functional clarity begins to blur.

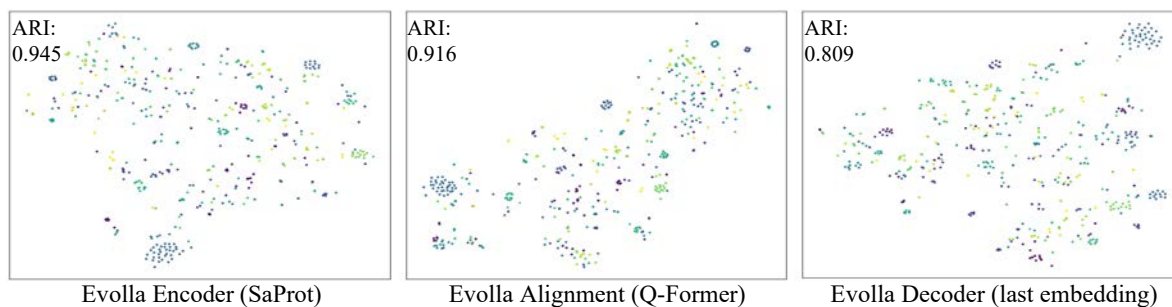


Figure 3: Visualization of representation spaces at different stages within the Evolla-10B model.

**Takeaway:** The degradation of functional representation stems not from the initial protein encoding, but from the subsequent semantic alignment to the language model.

## 5.4 Collapse on Novel Protein Families

A critical limitation of many large-scale models is their tendency to overfit to training data, leading to poor generalization on novel examples. We adopted the evaluation protocol from Evolla [38], which partitions the test set into three subsets based on sequence identity to the training set: Easy, Medium, and Hard. The division of these subsets is described in Appendix B.

The results, illustrated in Figure 4, reveal a dramatic divergence in generalization capability. Evolla’s performance exhibits a steep, monotonic decline as the data hardness increases. It performs well on the Easy subset with an LLM score of 81.9, where it can likely rely on memorized patterns from similar training sequences. The performance collapse of about 30% from Easy to Hard is a classic symptom of poor generalization. In stark contrast, our context-driven method demonstrates remarkable robustness. Its performance remains consistently high across all levels of difficulty. The performance is virtually unaffected by the novelty of the protein sequence. This stability stems from the fact that our approach does not rely on interpreting the raw sequence itself. Instead, it leverages high-level knowledge that are inherently designed to generalize well.

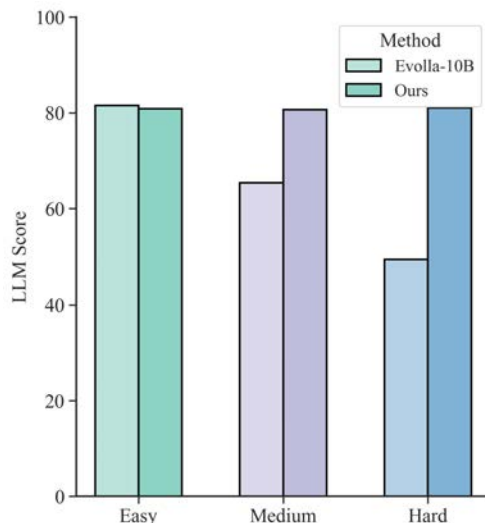


Figure 4: Comparison of Evolla-10B and our approach across the easy, medium, and hard subsets.

## 5.5 Degrading Phenomenon Across Time

We curated a dataset by randomly sampling about 100 proteins for each year from 1995 to 2024 based on the first publication year. The relationship between a protein’s first publication year and the models’ LLM-Scores is visualized in Figure 5.

For this analysis, our context-driven approach employed DeepSeek-V3 [24] as its base LLM to ensure a fair comparison against models with similar training data cut-off dates. **Our context-driven approach (a)**, while maintaining the highest overall performance, exhibits a slight negative trend over time due to the diminishing availability of rich, homologous information in the knowledge bases. For very recent proteins, homology-based tools like BLAST find fewer well-characterized relatives, leading to a sparser context and thus slightly less precise answers. **The sequence-as-modality model, Evolla (b)**, displays a much more pronounced degradation. Its performance on well-studied proteins from the 1990s and early 2000s is strong, but it deteriorates significantly for proteins discovered in the last decade. It is crucial to note that Evolla’s training data, sourced from Swiss-Prot Release (202303), has a temporal bias. Therefore, part of this decline can be attributed to its lack of exposure to the most recent protein data. However, this training bias alone does not fully account for the steepness of the collapse. The trend suggests a deeper issue: Evolla’s encoder appears to rely heavily on the dense web of evolutionary information available for older, larger protein families. When faced with recent, potentially more unique proteins that lack this deep evolutionary context—a problem exacerbated by its training data cutoff—the encoder’s ability to generate meaningful representations weakens considerably. **The sequence-as-language model, Intern-S1 (c)**, shows a performance profile that is almost entirely flat and consistently low across the entire 30-year period. This lack of temporal trend, combined with its overall poor performance, indicates a fundamental failure to extract meaningful biological signals from the raw sequence.

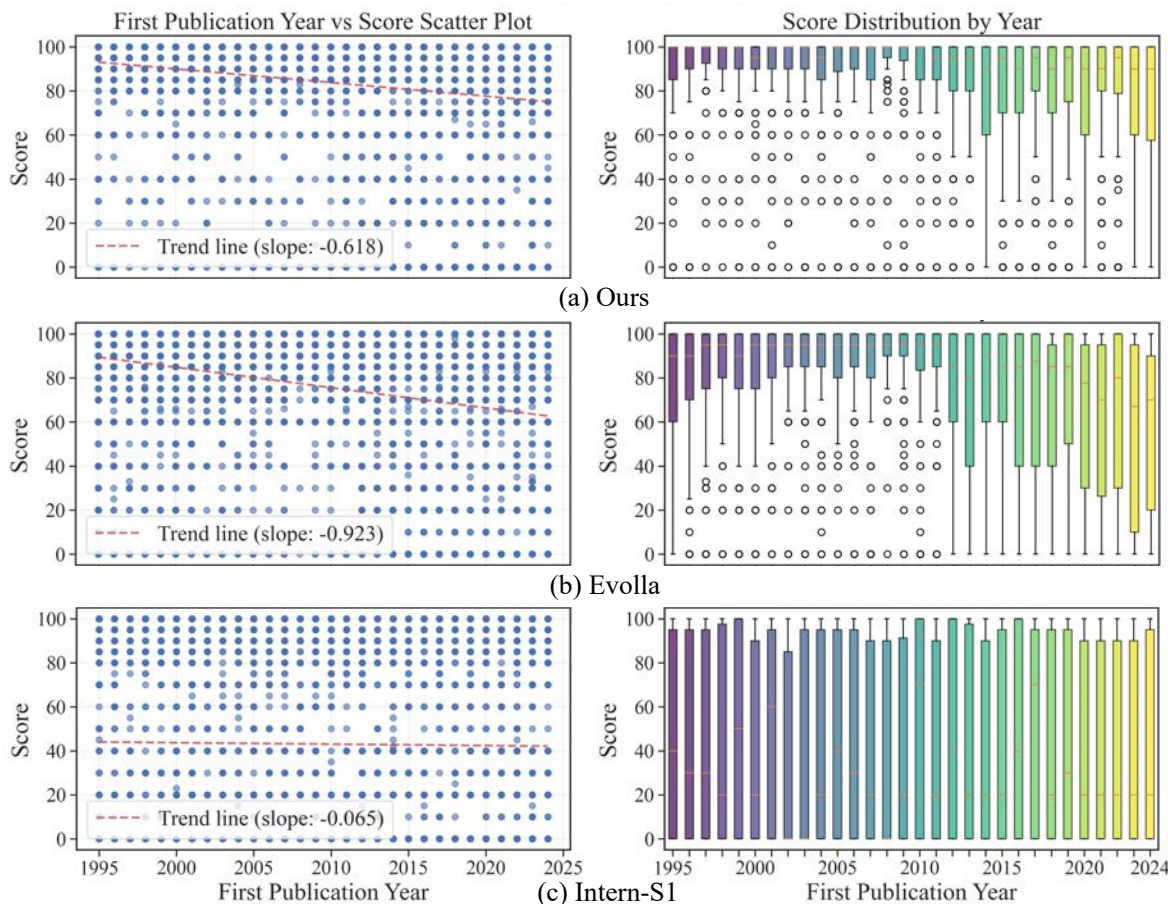


Figure 5: Analysis of model performance versus protein’s first publication year.

**Takeaway:** Our context-driven approach demonstrates superior generalization: (i) *Robustness to sequence novelty*: Unlike Sci-LLMs which suffer collapsing on proteins dissimilar to training data, our context maintains high accuracy regardless of sequence identity. (ii) *Temporal stability*: Our approach’s performance degrades far more gracefully over time on recently discovered proteins compared to other paradigms.

The above dual robustness confirms that reasoning over stable, high-level knowledge is a more robust foundation for AI in biology than relying on the difficult task of raw sequence interpretation.

## 6 Conclusion and Limitation

In this work, we confronted a fundamental challenge at the heart of modern Sci-LLMs: the tokenization dilemma. We demonstrated that current paradigms, whether treating biomolecular sequences as a specialized language or as a distinct modality, are fundamentally handicapped by issues of weak representation and semantic misalignment. Our central contribution is the validation of a third paradigm that resolves this dilemma. By shifting the focus from low-level sequence interpretation to high-level knowledge synthesis, our context-driven approach entirely circumvents the tokenization problem, as illustrated in the conceptual landscape of Figure 6. Our approach is computationally efficient, as it leverages generalist LLMs without the retraining required by domain-specific Sci-LLMs.

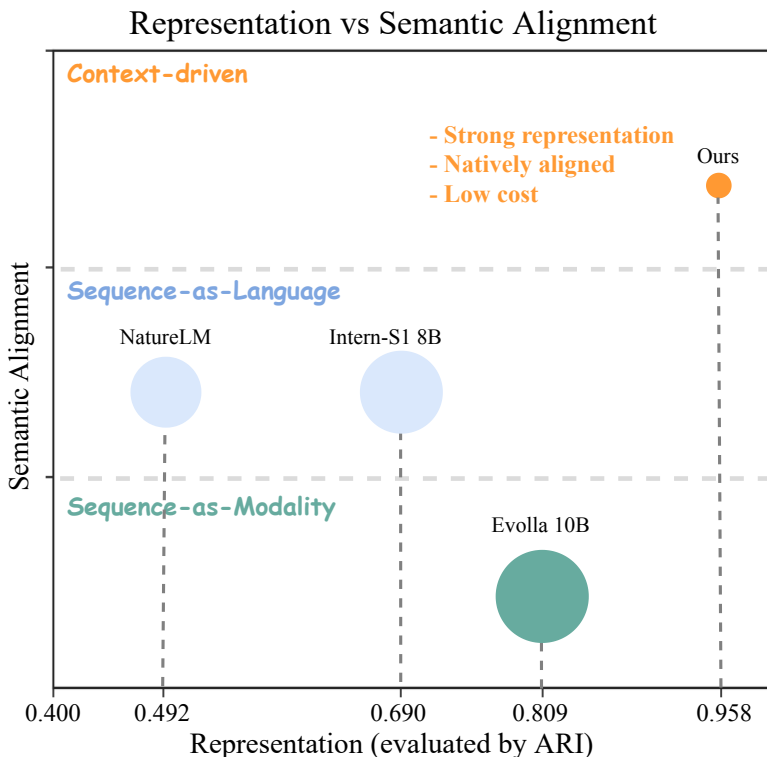


Figure 6: The trade-off landscape of representation vs. semantic alignment. The x-axis quantifies the quality of the biological representation (measured by ARI), while the y-axis conceptually represents the degree of semantic alignment with natural language. The area of each circle is proportional to the computational cost, with larger circles indicating higher computational expenses.

While our findings are compelling, we acknowledge several limitations. For truly novel orphan proteins from unexplored regions of the protein universe, our method’s performance may be constrained. Furthermore, our current analysis has primarily focused on proteins; although we provide some preliminary exploration in Appendix G, a more comprehensive treatment remains for future work.

## References

- [1] Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein’s function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10757–10765, 2024.
- [2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- [3] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- [4] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.
- [5] Lei Bai, Zhongrui Cai, Maosong Cao, Wei Han Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.
- [6] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O’Donovan, Isabelle Phan, et al. The swiss-prot protein knowledge-base and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.
- [7] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.
- [8] Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.
- [9] 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. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.
- [10] Bernardo P de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, et al. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, pages 1–14, 2025.
- [11] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127, 2021.
- [12] Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimier, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. Bioreason: Incentivizing multimodal biological reasoning within a dna-llm model, 2025. URL <https://arxiv.org/abs/2505.23579>.
- [13] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [14] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.
- [15] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.
- [16] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.

- [17] Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, et al. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*, 2025.
- [18] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025.
- [19] Yunha Hwang, Andre L Cornman, Elizabeth H Kellogg, Sergey Ovchinnikov, and Peter R Girguis. Genomic language model predicts protein co-regulation and function. *Nature communications*, 15(1):2880, 2024.
- [20] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- [21] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.
- [22] Anuj Karpatne, Aryan Deshwal, Xiaowei Jia, Wei Ding, Michael Steinbach, Aidong Zhang, and Vipin Kumar. Ai-enabled scientific revolution in the age of generative ai: second nsf workshop report. *npj Artificial Intelligence*, 1(1):18, 2025.
- [23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- [24] Aixiu Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [25] Nuowei Liu, Changzhi Sun, Tao Ji, Junfeng Tian, Jianxin Tang, Yuanbin Wu, and Man Lan. Evollama: Enhancing llms’ understanding of proteins via multimodal structure and sequence representations. *arXiv preprint arXiv:2412.11618*, 2024.
- [26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [27] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.
- [28] 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. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):ead9336, 2024.
- [29] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pages 2020–12, 2020.
- [30] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.
- [31] Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, pages 2024–05, 2024.
- [32] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- [33] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, pages 1–9, 2025.

- [34] Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, et al. Naturelm: Deciphering the language of nature for scientific discovery. *arXiv e-prints*, pages arXiv–2502, 2025.
- [35] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [36] Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.
- [37] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.
- [38] Xibin Zhou, Chenchen Han, Yingqi Zhang, Jin Su, Kai Zhuang, Shiyu Jiang, Zichen Yuan, Wei Zheng, Fengyuan Dai, Yuyang Zhou, et al. Decoding the molecular language of proteins with evolla. *bioRxiv*, pages 2025–01, 2025.
- [39] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024.