

000 001 002 003 004 005 006 007 008 009 010 011 LOST IN TOKENIZATION: CONTEXT AS THE KEY 012 TO UNLOCKING BIOMOLECULAR UNDERSTANDING IN 013 SCIENTIFIC LLMs 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037

Anonymous authors

Paper under double-blind review

ABSTRACT

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

1 INTRODUCTION

038 The convergence of artificial intelligence and the life sciences has given rise to a new class of pow-
039 erful tools: Scientific Large Language Models (Sci-LLMs). Built on Transformer architectures (e.g.
040 BERT, GPT) that have revolutionized natural language processing (Ferruz et al., 2022), these models
041 hold immense promise for accelerating biological discovery (Karpatne et al., 2025). From predicting
042 protein function (Brandes et al., 2023) to designing novel therapeutics (Hie et al., 2024), Sci-LLMs
043 such as Intern-S1 (Bai et al., 2025), Evolla (Zhou et al., 2025), and NatureLM (Xia et al., 2025)
044 are being developed to interpret the complex “language of life” encoded in DNA, RNA, and protein
045 sequences (Taylor et al., 2022). Early efforts have demonstrated their potential, sparking visions
046 of an AI-driven future for scientific research. This burgeoning field has largely coalesced around
047 two primary strategies for integrating biomolecular data (Hwang et al., 2024). The first “*sequence-*
048 *as-language*” approach treats sequences as a specialized form of language, extending the model’s
049 vocabulary to include individual amino acids or nucleotides and pre-training it on vast corpora of se-
050 quence and text data. The second “*sequence-as-modality*” approach, inspired by multimodal learn-
051 ing, treats sequences as a distinct modality, employing a specialized encoder (e.g., a pre-trained
052 biological foundation model like ESM (Lin et al., 2023) and Evo (Fallahpour et al., 2025)) to gen-
053 erate 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 (Abdine et al., 2024; Liu et al., 2024b; de Almeida et al., 2025).

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 (Rao et al., 2020; Brandes et al., 2023). 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 (Ferruz et al., 2022). 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 (Hu et al., 2025). 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.

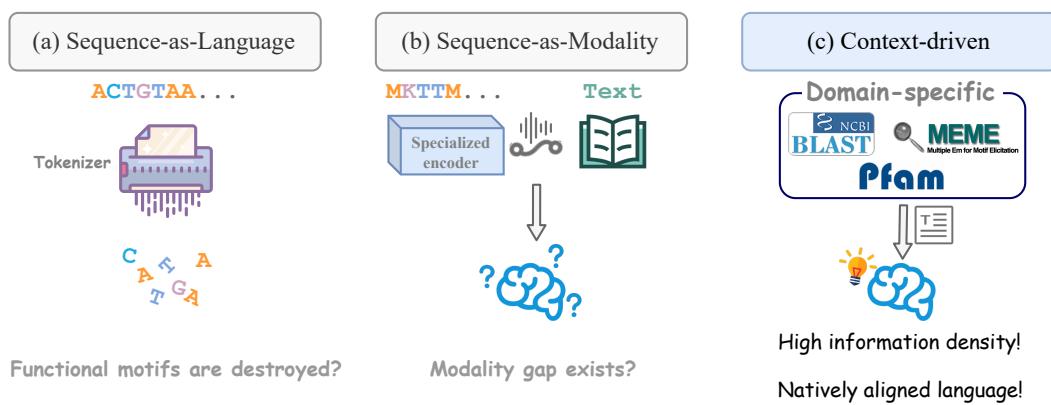


Figure 1: Paradigms for integrating biomolecular sequences into Sci-LLMs. (a) The sequence-as-language approach, tokenization fragments sequences into atomic symbols, potentially destroying functional motifs. (b) The sequence-as-modality approach preserves structure via specialized encoders but suffers 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.

108

2 RELATED WORK

109

110 2.1 FOUNDATION MODELS IN BIOLOGICAL REPRESENTATION

111 Foundation models for biological sequences have made rapid strides in representation learning. In
 112 the protein domain, large language models like ProtBERT (Elnaggar et al., 2021) and the ESM
 113 series (Lin et al., 2023; Hayes et al., 2025) are trained on massive sequence corpora, capturing signals
 114 of evolutionary conservation, structural motifs, and residue co-variation that enable downstream
 115 generalization. On the nucleotide side, models such as DNABERT (Ji et al., 2021) and the more
 116 recent Nucleotide Transformer (Dalla-Torre et al., 2025) apply k -mer tokenization or other sub-
 117 word strategies to genome-scale data, achieving high accuracy in identifying promoters, splice sites,
 118 and transcription factor binding locations. Multi-species genome models like DNABERT-2 (Zhou
 119 et al., 2024) further improve efficiency by replacing k -mers with Byte-Pair Encoding to accom-
 120 modate longer input sequences. Meanwhile, specialized transformer architectures have extended
 121 context lengths to capture distal regulatory interactions and boost gene expression prediction (Avsec
 122 et al., 2021; Nguyen et al., 2023; 2024). Despite their powerful representational capacity, these
 123 bio-sequence foundation models largely act as “black boxes”. Their internal embeddings are high-
 124 dimensional and not straightforwardly mapped to human-interpretable biological units like motifs,
 125 domains, or pathways, making it difficult to extract mechanistic insight.

126

127 2.2 SCIENTIFIC LARGE LANGUAGE MODELS

128 Large language models tailored to scientific domains (Sci-LLMs) have rapidly advanced, extending
 129 the success of general LLMs into tasks like protein or molecule design, genomic analysis, and sci-
 130 entific reasoning. Galactica (Taylor et al., 2022), a 120-billion-parameter model trained on a corpus of
 131 papers and knowledge bases, was introduced to store and reason over scientific knowledge. Domain-
 132 focused sequence models have also emerged: NatureLM (Xia et al., 2025), for example, is a unified
 133 sequence-based model pre-trained across proteins, nucleic acids and small molecules. Likewise,
 134 Intern-S1 (Bai et al., 2025) is a recent large multimodal MoE model (28B activated parameters)
 135 with specialized tokenization and encoders for different scientific modalities. In this work, we focus
 136 on biomolecular understanding as a representative scientific challenge: information is inherently
 137 encoded in sequences (genes or proteins), which can be expressed in textual form or as a distinct
 138 modality, making it an ideal testbed for probing how well Sci-LLMs integrate domain knowledge
 139 and whether they truly understand biological sequences.

140

141 2.3 EXISTING STRATEGIES IN BRIDGING SEQUENCES AND LANGUAGE

142 Sci-LLMs have adopted several strategies to bridge low-level biomolecular sequences with higher-
 143 level reasoning and knowledge. One common approach is treating sequences as a specialized lan-
 144 guage. Models like NatureLM (Xia et al., 2025) and Intern-S1 (Bai et al., 2025) ingest raw or
 145 tokenized sequences directly as input, training on vast datasets of sequences annotated with text so
 146 that the model learns joint representations. Another emerging strategy is treating sequences as a
 147 separate modality. For example, EvOLLaMA (Liu et al., 2024b) incorporates a protein structure en-
 148 coder and a sequence encoder alongside an LLM to enable multimodal protein question-answering,
 149 and Evolla (Zhou et al., 2025) employs SaProt Su et al. (2024a) as the structure encoder. BioRea-
 150 son (Fallahpour et al., 2025) similarly couples a frozen DNA foundation model Evo (Nguyen et al.,
 151 2024) with a language model Qwen3 (Yang et al., 2025), so that genomic sequences are converted
 152 into contextual embeddings which the LLM can reason over in natural language. A third line of
 153 work explores agent-based or tool-augmented approaches. Rather than having a single model di-
 154 rectly analyze sequences, the LLM is equipped with the ability to call external tools or databases as
 155 needed. Notable examples include GeneAgent (Wang et al., 2025), which self-verifies for gene-set
 156 analysis using domain databases, and ChemCrow (Bran et al., 2023), which uses an agent to plan
 157 multi-step chemistry tasks by invoking a suite of expert tools. While all these strategies have pushed
 158 the frontier of scientific AI (Huang et al., 2025), it remains unclear how much of the success in
 159 Sci-LLMs comes from genuine reasoning over raw sequences. In this work, we adopt a deliberately
 160 context-driven baseline—providing the model with only high-level, structured annotations of the
 161 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.

162

3 PRELIMINARIES

164

3.1 THE BIOMOLECULAR UNDERSTANDING TASK

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

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

$$172 \quad a = \mathcal{M}(s, q; \theta) \quad (1)$$

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

177

3.2 SEQUENCE-AS-LANGUAGE

179 This approach, utilized by models such as NatureLM (Xia et al., 2025) and Intern-S1 (Bai et al.,
 180 2025), treats a biomolecular sequence as a specialized string of text. Let T_{seq} be a tokenizer that
 181 maps a sequence s into a series of tokens from a biological vocabulary, V_{bio} , and let T_{text} be a
 182 standard tokenizer for a natural language question q with vocabulary V_{text} . The model operates
 183 on an extended vocabulary $V_{ext} = V_{text} \cup V_{bio}$. The input to the LLM, X_{input} , is formed by the
 184 concatenation of the tokenized question and sequence:

$$185 \quad X_{input} = [T_{text}(q); T_{seq}(s)] \quad (2)$$

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

$$188 \quad P(a|s, q) = \prod_{k=1}^{|a|} P(a_k|a_{<k}, X_{input}; \theta) \quad (3)$$

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

195

3.3 SEQUENCE-AS-MODALITY

197 Inspired by successes in vision-language modeling, this paradigm—employed by models like
 198 Evolla (Zhou et al., 2025) and BioReason (Fallahpour et al., 2025)—treats the biomolecular se-
 199 quence as a distinct, non-textual modality. A specialized, pre-trained biomolecular encoder,
 200 $\mathcal{E}_{bio} : \mathcal{S} \rightarrow \mathbb{R}^{L \times d}$, first transforms the sequence s into a sequence of rich, contextualized em-
 201 beddings. An alignment module, \mathcal{A}_{align} , then projects these biological embeddings into the LLM’s
 202 semantic space, creating an aligned sequence representation $E_{aligned_seq} \in \mathbb{R}^{K \times d}$. The final input to
 203 the LLM is a structured combination of the embedded text and the aligned sequence embeddings:

$$204 \quad X_{input} = [T_{text}(q); E_{aligned_seq}] \quad (4)$$

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

211

4 THE CONTEXT-DRIVEN APPROACH

214 In this work, we propose and investigate a third paradigm that circumvents the tokenization dilemma
 215 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 $\mathcal{C} : \mathcal{S} \rightarrow \mathcal{T}_{\text{context}}$, where $\mathcal{T}_{\text{context}}$ 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 :

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

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

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

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 (Jones et al., 2014) is used to identify conserved domains and motifs based on the sequence’s intrinsic features, while BLASTp (Altschul et al., 1990) retrieves annotations from close homologs in the Swiss-Prot database (Boeckmann et al., 2003). For novel orphan sequences lacking hits from these tools, we use the tri-modal retrieval model ProTrek (Su et al., 2024b) 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.

270 5 THE TOKENIZATION DILEMMA IN PRACTICE
271272 5.1 THE PRIMACY OF CONTEXT OVER SEQUENCE
273

274 Following a standardized protocol inspired by Evolla (Zhou et al., 2025), our benchmark focuses on
275 three fundamental aspects of protein biology: molecular function, metabolic pathway involvement,
276 and subcellular localization. For each protein in our test set, we generated queries corresponding to
277 these categories (e.g., “What is the function of this protein?”). To ensure a set of factually grounded
278 and verifiable ground truths, a question was only included if its corresponding annotation field was
279 explicitly present in the source database entry, from which the answer was directly excerpted. Per-
280 formance was quantified using an automated pipeline, leveraging a general-purpose LLM as an
281 expert judge, a metric we term the LLM-Score. A detailed description of the dataset construction,
282 evaluation protocol, and prompt design is provided in Appendices B and C. We evaluate the per-
283 formance of both specialized Sci-LLMs and leading general-purpose LLMs across three distinct
284 input configurations: (i) Sequence-Only, where the model receives only the raw protein sequence;
285 (ii) Sequence + Context, a combined input; (iii) Context-Only, where the model receives only the
286 high-level context. The results are presented in Table 1.

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

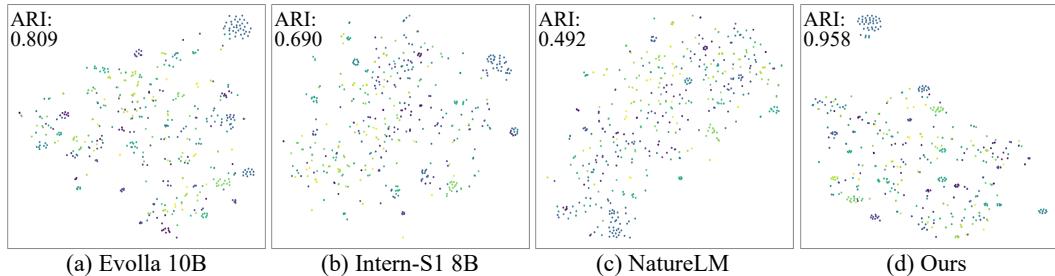
291 Model	292 Sequence	293 Context	294 Func.	295 Path.	296 Sub. Loc.	297 All
Specialized Sci-LLMs						
298 Intern-S1	✓		20.57	26.56	69.75	43.33
299 Intern-S1	✓	✓	74.18	98.85	93.00	84.03
300 Intern-S1		✓	76.22	97.60	95.60	<u>86.15</u>
301 Evolla	✓		40.23	72.71	79.76	59.93
302 Evolla	✓	✓	57.46	84.69	83.05	70.53
303 Evolla		✓	65.77	83.33	81.88	<u>74.02</u>
304 NatureLM	✓		3.58	5.52	10.45	6.82
305 NatureLM	✓	✓	42.33	64.25	32.30	38.86
306 NatureLM		✓	44.77	51.35	32.51	<u>39.50</u>
General LLMs						
307 Deepseek-v3	✓		10.98	24.54	74.72	40.77
308 Deepseek-v3	✓	✓	77.40	91.35	94.75	<u>86.03</u>
309 Deepseek-v3		✓	75.79	93.96	93.65	84.99
310 Gemini2.5 Pro	✓		10.40	13.85	77.58	41.25
311 Gemini2.5 Pro	✓	✓	79.12	94.17	94.65	86.98
312 Gemini2.5 Pro		✓	79.17	98.65	94.56	87.19
313 GPT-5	✓		19.64	17.08	64.15	39.83
314 GPT-5	✓	✓	79.89	89.48	71.30	<u>76.45</u>
315 GPT-5		✓	77.25	85.73	73.05	75.76
316 Qwen3-235B-A22B	✓		<u>13.67</u>	<u>19.90</u>	<u>37.17</u>	<u>39.51</u>
317 Qwen3-235B-A22B	✓	✓	<u>76.62</u>	<u>96.35</u>	<u>94.78</u>	<u>85.90</u>
318 Qwen3-235B-A22B		✓	<u>75.63</u>	<u>92.19</u>	<u>94.28</u>	<u>84.99</u>

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

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

324 5.2 DECONSTRUCTING THE DILEMMA I: THE WEAK REPRESENTATION
325

326 We visualize the embeddings of the outputs, where ground-truth classes were established by clustering
327 homologous proteins using MMseqs2 at a 50% sequence identity threshold. For each model,
328 we extracted the final-layer embeddings for their outputs. We employed t-SNE (Maaten & Hinton,
329 2008) to project them into a 2D space. The quality of the resulting functional separation was
330 then quantified by performing clustering on the high-dimensional embeddings and calculating the
331 Adjusted Rand Index (ARI) against the MMseqs2 ground-truth clusters. For our context-driven
332 approach, we generated embeddings from the structured context itself using the text embedding model
333 Qwen-embedding (Zhang et al., 2025). The results are visualized in Figure 2.

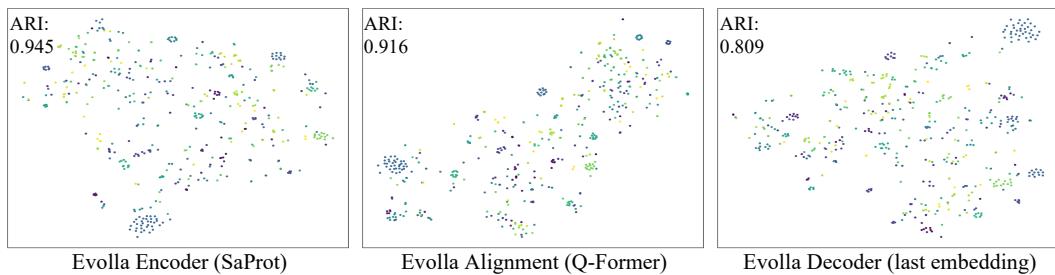
343 Figure 2: The visualization of representation spaces.
344

345 **Takeaway:** Simple context provides a vastly superior functional representation of proteins
346 compared to both sequence-to-language/modality strategies.
347

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

355 5.3 DECONSTRUCTING THE DILEMMA II: THE SEMANTIC MISALIGNMENT
356

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

374 Figure 3: Visualization of representation spaces at different stages within the Evolla-10B model.
375

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

378 5.4 COLLAPSE ON NOVEL PROTEIN FAMILIES
379

380 A critical limitation of many large-scale models is their tendency to overfit to training data, leading
381 to poor generalization on novel examples. We adopted the evaluation protocol from Evolla (Zhou
382 et al., 2025), which partitions the test set into three subsets based on sequence identity to the training
383 set: Easy, Medium, and Hard. The division of these subsets is described in Appendix B.

384 The results, illustrated in Figure 4, reveal a dramatic divergence in generalization capability.
385 Evolla’s performance exhibits a steep, monotonic decline as the data hardness increases. It
386 performs well on the Easy subset with an LLM score of 81.9, where it can likely rely on memo-
387 rized patterns from similar training sequences.
388 The performance collapse of about 30% from
389 Easy to Hard is a classic symptom of poor
390 generalization. In stark contrast, our context-
391 driven method demonstrates remarkable robust-
392 ness. Its performance remains consistently high
393 across all levels of difficulty. The performance
394 is virtually unaffected by the novelty of the
395 protein sequence. This stability stems from the
396 fact that our approach does not rely on interpret-
397 ing the raw sequence itself. Instead, it lever-
398 ages high-level knowledge that are inherently
399 designed to generalize well.
400

402 5.5 DEGRADING PHENOMENON ACROSS TIME
403

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

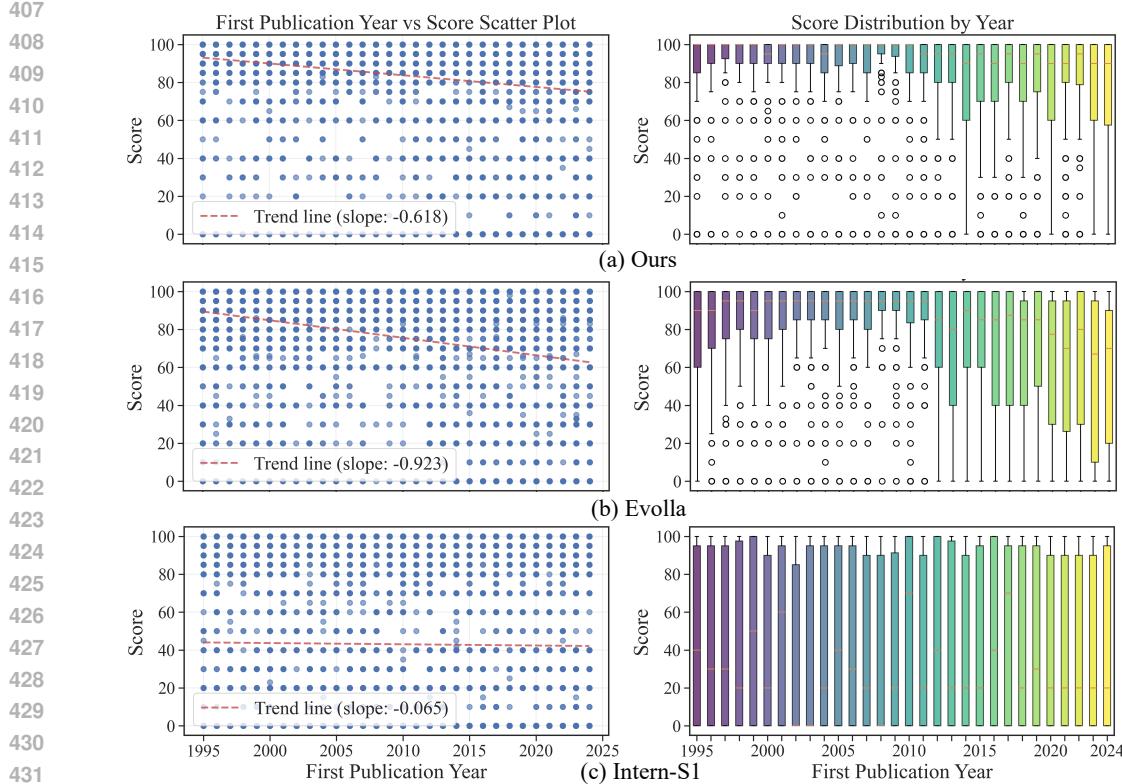


Figure 4: Comparison of Evolla-10B and our approach across the easy, medium, and hard subsets.

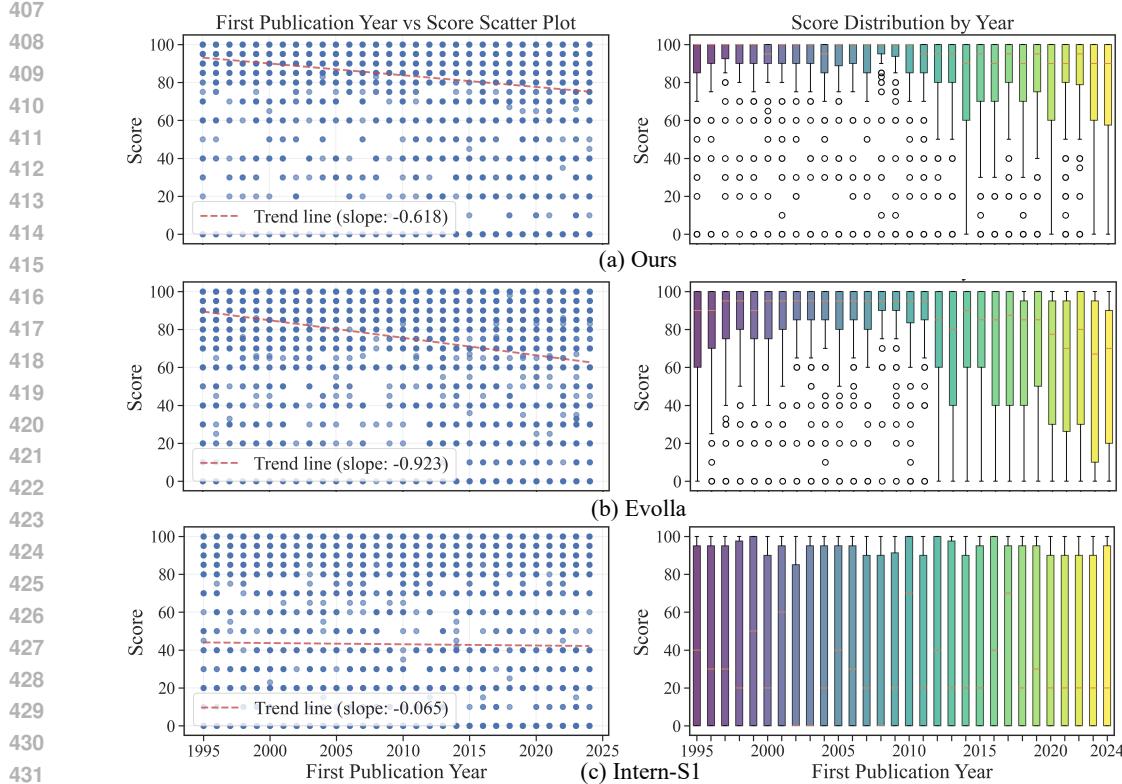


Figure 5: Analysis of model performance versus protein’s first publication year.

For this analysis, our context-driven approach employed DeepSeek-V3 (Liu et al., 2024a) 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.

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.

5.6 QUANTITATIVE ANALYSIS OF REAL-WORLD EFFICIENCY

We conducted a detailed, three-way analysis to quantify the trade-offs between computational cost, time, and performance. We compare: (i) A general LLM baseline Deepseek-v3, which feeds the raw sequence to the DeepSeek-V3 API and yields a performance score of **40.77**, (ii) A specialized Sci-LLM Evolla, which requires a high-end GPU and achieves a performance score of **59.93**, and (iii) our context-driven method, which uses bioinformatics tools on a CPU plus the DeepSeek-V3 API to achieve a state-of-the-art performance score of **84.99**. We conducted the analysis across both single-sequence and batch-processing scenarios. The results, based on AWS on-demand pricing, are summarized in Table 2, with the cost estimation detailed in Appendix M.

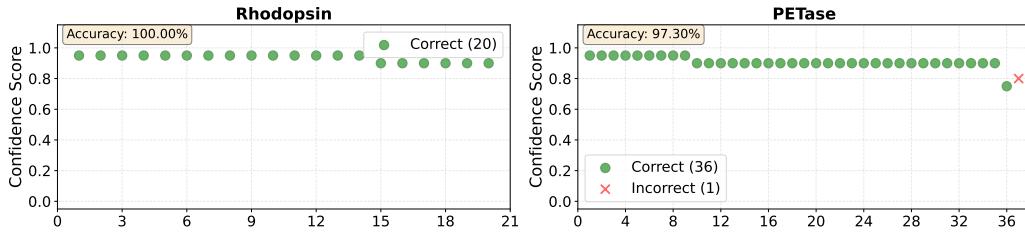
For individual queries, our method is not only dramatically more effective but also approximately **23 times cheaper** and **1.3 times faster** than the specialized end-to-end model. The true efficiency of our pipeline is most evident in high-throughput research. In a realistic, large-scale scenario, our method is nearly **30 times cheaper** and, critically, **154 times faster** than Evolla on a per-sequence basis. While the direct API baseline is cheaper, its poor performance renders it unsuitable for scientific applications. In summary, the small cost of running established bioinformatics tools is a highly effective investment, unlocking a $> 2 \times$ performance gain over the LLM baseline while simultaneously outperforming and costing less than specialized end-to-end models, especially at scale.

Table 2: Comparative analysis of inference efficiency.

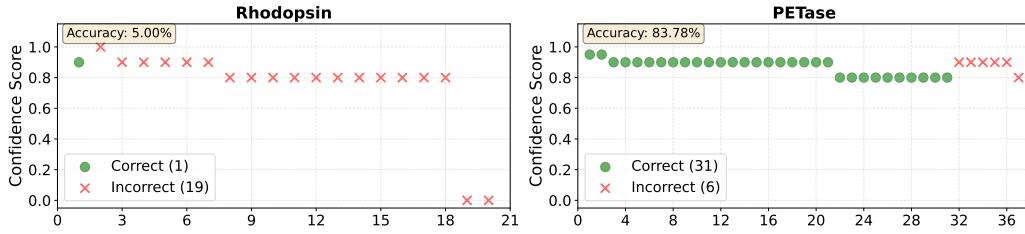
Method	Mode	Input to LLM	Avg. Time	Avg. Cost
Deepseek-v3	Single	Raw sequence	~30s	\$0.0005
Evolla	Single	Raw sequence	~90s	\$0.0690
Our Method	Single	Context	~70s	\$0.0030
Evolla	Batch	Raw sequence	~20s	\$0.0152
Our Method	Batch	Context	~0.13s	\$0.0005

486 5.7 WET-LAB VALIDATION ON NOVEL SEQUENCES
487

488 We curated a set of novel functional protein sequences obtained from wet-lab experiments. They
489 were unpublished at the time of our analysis and absent from major databases, including Swiss-Prot,
490 thereby representing a true test of performance on unseen data. The task was formulated as a binary
491 classification problem for two distinct protein families: Rhodopsin and PETase. For each sequence,
492 the LLM was prompted to predict its classification. As shown in Figure 6, our method achieves
493 100% accuracy for Rhodopsin and 97.3% accuracy for PETase. While Evolla (Figure 7) attains a
494 reasonable 80.0% accuracy on Rhodopsin, it fails catastrophically on PETase. It may be caused by its
495 training data bias. A detailed description of the wet-lab validation is shown in Appendix N.



503 Figure 6: Sample-level performance of our context-driven method + Deepseek-v3.

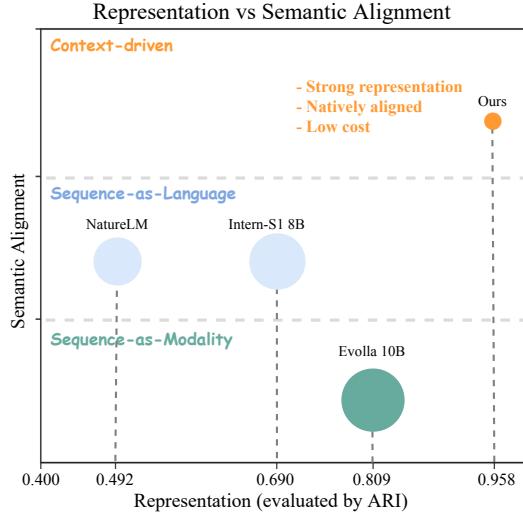


511 Figure 7: Sample-level performance of Evolla.

512 6 CONCLUSION AND LIMITATION

513 In this work, we confronted a fundamental challenge at the heart of modern Sci-LLMs: the
514 tokenization dilemma. We demonstrated that
515 current paradigms, whether treating biomolecular
516 sequences as a specialized language or as
517 a distinct modality, are fundamentally handicapped
518 by issues of weak representation and
519 semantic misalignment. Our central contribution
520 is the validation of a third paradigm that
521 resolves this dilemma. By shifting the focus
522 from low-level sequence interpretation to high-
523 level knowledge synthesis, our context-driven
524 approach entirely circumvents the tokenization
525 problem, as illustrated in the conceptual land-
526 scape of Figure 8. Notably, our approach is also
527 computationally efficient, as it leverages generalist
528 LLMs without the costly retraining re-
529 quired by domain-specific Sci-LLMs.

530 While our findings are compelling, we ac-
531 knowledge several limitations. For truly novel
532 orphan proteins from unexplored regions of
533 the protein universe, our method’s performance
534 may be constrained. Furthermore, our cur-
535 rent analysis has primarily focused on proteins;
536 although we provide some preliminary explo-
537 ration in Appendix G, a more comprehensive
538 treatment remains open for future research.



539 Figure 8: The trade-off landscape of representation vs. semantic alignment. The x-axis quantifies
540 the quality of the biological representation (mea-
541 sured by ARI), while the y-axis conceptually rep-
542 resents the degree of semantic alignment with nat-
543 ural language. The area of each circle is pro-
544 portional to the computational cost, with larger
545 circles indicating higher computational expenses.

540 REFERENCES

- 541
- 542 Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text:
543 Multimodal protein’s function generation with gnns and transformers. In *Proceedings of the AAAI
544 Conference on Artificial Intelligence*, volume 38, pp. 10757–10765, 2024.
- 545 Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local
546 alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- 547
- 548 Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael
549 Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool
550 for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- 551
- 552 Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska,
553 Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene
554 expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18
555 (10):1196–1203, 2021.
- 556
- 557 Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojong Chen, Kai Chen,
558 Pengcheng Chen, Ying Chen, Yongkang Chen, et al. Intern-s1: A scientific multimodal foun-
559 dation model. *arXiv preprint arXiv:2508.15763*, 2025.
- 560
- 561 Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elis-
562 abeth Gasteiger, Maria J Martin, Karine Michoud, Claire O’Donovan, Isabelle Phan, et al. The
563 swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31
(1):365–370, 2003.
- 564
- 565 Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe
566 Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint
567 arXiv:2304.05376*, 2023.
- 568
- 569 Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-
570 wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*,
55(9):1512–1522, 2023.
- 571
- 572 Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk
573 Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan
574 Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for
575 human genomics. *Nature Methods*, 22(2):287–297, 2025.
- 576
- 577 Bernardo P de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer,
578 Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, et al. A
579 multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, pp.
1–14, 2025.
- 580
- 581 Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones,
582 Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward un-
583 derstanding the language of life through self-supervised learning. *IEEE transactions on pattern
584 analysis and machine intelligence*, 44(10):7112–7127, 2021.
- 585
- 586 Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah,
587 Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and Bo Wang. Bioreason:
588 Incentivizing multimodal biological reasoning within a dna-llm model, 2025. URL <https://arxiv.org/abs/2505.23579>.
- 589
- 590 Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and
591 Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language
592 models. In *The Twelfth International Conference on Learning Representations*, 2024.
- 593
- Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model
for protein design. *Nature communications*, 13(1):4348, 2022.

- 594 Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert
 595 Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years
 596 of evolution with a language model. *Science*, 387(6736):850–858, 2025.
- 597
- 598 Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng
 599 Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from
 600 general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.
- 601
- 602 Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang
 603 Zhuang, Jiaqi Liu, Yingzhou Lu, et al. A survey of scientific large language models: From data
 604 foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*, 2025.
- 605
- 606 Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li,
 607 Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*,
 608 2025.
- 609
- 610 Yunha Hwang, Andre L Cornman, Elizabeth H Kellogg, Sergey Ovchinnikov, and Peter R Girguis.
 611 Genomic language model predicts protein co-regulation and function. *Nature communications*,
 612 15(1):2880, 2024.
- 613
- 614 Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional
 615 encoder representations from transformers model for dna-language in genome. *Bioinformatics*,
 616 37(15):2112–2120, 2021.
- 617
- 618 Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish
 619 McWilliam, John Maslen, Alex Mitchell, Gift Nuka, et al. Interproscan 5: genome-scale protein
 620 function classification. *Bioinformatics*, 30(9):1236–1240, 2014.
- 621
- 622 Anuj Karpatne, Aryan Deshwal, Xiaowei Jia, Wei Ding, Michael Steinbach, Aidong Zhang, and
 623 Vipin Kumar. Ai-enabled scientific revolution in the age of generative ai: second nsf workshop
 624 report. *npj Artificial Intelligence*, 1(1):18, 2025.
- 625
- 626 Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin,
 627 Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level
 628 protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
- 629
- 630 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 631 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
 632 *arXiv:2412.19437*, 2024a.
- 633
- 634 Nuowei Liu, Changzhi Sun, Tao Ji, Junfeng Tian, Jianxin Tang, Yuanbin Wu, and Man Lan. Evol-
 635 lama: Enhancing llms’ understanding of proteins via multimodal structure and sequence repre-
 636 sentations. *arXiv preprint arXiv:2412.11618*, 2024b.
- 637
- 638 Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine
 639 learning research*, 9(Nov):2579–2605, 2008.
- 640
- 641 Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes,
 642 Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range
 643 genomic sequence modeling at single nucleotide resolution. *Advances in neural information
 644 processing systems*, 36:43177–43201, 2023.
- 645
- 646 Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruba Katrekar, David B Li, Liam J
 647 Bartie, Armin W Thomas, Samuel H King, Garyk Bixi, et al. Sequence modeling and design
 648 from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024.
- 649
- 650 Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer
 651 protein language models are unsupervised structure learners. *Biorxiv*, pp. 2020–12, 2020.
- 652
- 653 Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein
 654 language modeling with structure-aware vocabulary. In *The Twelfth International Conference on
 655 Learning Representations*, 2024a.

- 648 Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. Protrek: Navigating the protein universe through
 649 tri-modal contrastive learning. *bioRxiv*, pp. 2024–05, 2024b.
 650
- 651 Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia,
 652 Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for
 653 science. *arXiv preprint arXiv:2211.09085*, 2022.
- 654 Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day,
 655 Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for
 656 gene-set analysis using domain databases. *Nature Methods*, pp. 1–9, 2025.
 657
- 658 Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang,
 659 Zequn Liu, Yuan-Jyue Chen, et al. Naturelm: Deciphering the language of nature for scientific
 660 discovery. *arXiv e-prints*, pp. arXiv–2502, 2025.
- 661 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,
 662 Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint
 663 arXiv:2505.09388*, 2025.
- 664 Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme
 665 function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.
 666
- 667 Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie,
 668 An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and
 669 reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.
- 670 Xibin Zhou, Chenchen Han, Yingqi Zhang, Jin Su, Kai Zhuang, Shiyu Jiang, Zichen Yuan, Wei
 671 Zheng, Fengyuan Dai, Yuyang Zhou, et al. Decoding the molecular language of proteins with
 672 evolla. *bioRxiv*, pp. 2025–01, 2025.
 673
- 674 Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. Dnabert-2:
 675 Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth Interna-*
 676 *tional Conference on Learning Representations*, 2024.
 677
 678
 679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701

702	CONTENTS	
703		
704		
705	1 Introduction	1
706		
707	2 Related Work	3
708	2.1 Foundation Models in Biological Representation	3
709	2.2 Scientific Large Language Models	3
710	2.3 Existing Strategies in Bridging Sequences and Language	3
711		
712		
713	3 Preliminaries	4
714	3.1 The Biomolecular Understanding Task	4
715	3.2 Sequence-as-Language	4
716	3.3 Sequence-as-Modality	4
717		
718		
719		
720	4 The Context-Driven Approach	4
721		
722	5 The Tokenization Dilemma in Practice	6
723	5.1 The Primacy of Context over Sequence	6
724	5.2 Deconstructing the Dilemma I: The Weak Representation	7
725	5.3 Deconstructing the Dilemma II: The Semantic Misalignment	7
726	5.4 Collapse on Novel Protein Families	8
727	5.5 Degrading Phenomenon Across Time	8
728	5.6 Quantitative Analysis of Real-World Efficiency	9
729	5.7 Wet-lab Validation on Novel Sequences	10
730		
731		
732		
733		
734	6 Conclusion and Limitation	10
735		
736		
737	A Context details	15
738	A.1 Context Generation	15
739	A.2 Context Construction	16
740	A.3 Context-based Inference	16
741		
742		
743	B Dataset details	16
744	B.1 Protein Dataset	16
745	B.2 DNA Dataset	18
746		
747		
748	C Evaluation Metric	18
749	C.1 LLM-Score for General Protein QA Tasks	18
750	C.2 Hierarchical Metrics for EC Number Prediction	19
751		
752		
753	D Quantitative Benchmark: EC Number Prediction	19
754		
755	E Ablation Study	20

756	F Impact of Semantic Alignment on Mutation Sensitivity	21
757		
758	G Generalizability Across Biomolecular Types	23
759		
760	G.1 Performance on Protein Function Predication	23
761	G.2 Performance on DNA Mutation Prediction	24
762		
763	H Independence from Clustering Metrics	24
764		
765	I QA examples	25
766		
767	I.1 Comparative Analysis on Protein A6LHQ9	25
768	I.1.1 Case 1: Context-Only (Success)	25
769	I.1.2 Case 2: Sequence-Only (Total Failure)	27
770	I.1.3 Case 3: Sequence + Context (Noise Effect)	28
771	I.2 Showcase of Our Method’s Success Cases	29
772	I.2.1 Example 1: Function Prediction	30
773	I.2.2 Example 2: Pathway Prediction	32
774	I.2.3 Example 3: Subcellular Location Prediction	33
775		
776		
777		
778		
779	J Limitations	34
780		
781	J.1 Wild-Type Protein Example	34
782	J.2 Mutated Protein Example	34
783	J.3 Summary of the Limitation	35
784		
785		
786	K Comparison of Model Performance on the Same Protein and Question	35
787		
788	K.1 Analysis of Model Differences	38
789		
790	L The Use of Large Language Models (LLMs)	38
791		
792	M Detailed Cost Estimation	39
793		
794	N Wet-lab Validation Samples	39
795		
796	N.1 Prompt	39
797	N.2 Sequence List	40
798		
799	A CONTEXT DETAILS	
800		

801 Our context-driven consists of three stages: (1) generation of a multi-source evidence profile, (2)
 802 hierarchical construction of a textual context, and (3) context-based inference by LLMs.
 803

804 A.1 CONTEXT GENERATION

805 For any given biomolecular sequence (we use proteins as the running example), we first generate a
 806 comprehensive functional profile using a deliberately crafted, multi-source toolchain.
 807

808 **Input:** A single protein FASTA sequence.
 809

Toolbox Execution:

- 810 1. **Feature-intrinsic domain analysis:** We first scan the input sequence with **Inter-**
 811 **ProScan** (Jones et al., 2014) to identify conserved domains and key motifs by integrating
 812 multiple signature libraries (e.g., Pfam, PROSITE, SMART). This step is an analysis grounded
 813 in *intrinsic* sequence features: even for a completely novel, unannotated protein, InterProScan
 814 can recognize known modular features. We extract textual descriptions of detected domains
 815 together with any directly linked Gene Ontology (GO) annotations (Ashburner et al., 2000) .
- 816 2. **Homology-based functional inference.** In parallel, we run **BLASTp** (Altschul et al., 1990)
 817 against a curated reference database (e.g., Swiss-Prot) to retrieve close homologs. We *transfer*
 818 GO annotations from the most similar sequences to the query. *Critically, we never use the*
 819 *query’s own (possibly unknown) labels.*
- 820 3. **Information integration.** We merge GO evidence obtained from InterProScan (feature-
 821 intrinsic) and from BLASTp (homology-based) to form a comprehensive functional profile.
- 822 4. **Fallback mechanism.** In rare cases where neither BLASTp nor InterProScan yields infor-
 823 mative signals (“orphan” sequences), we invoke **ProTrek** (Su et al., 2024b) to synthesize a
 824 concise, model-based textual description that serves as minimal context.

825 **A.2 CONTEXT CONSTRUCTION**

826 The raw outputs from these tools can be redundant, conflicting, or noisy, especially for novel pro-
 827 teins. A naive combination of all outputs is suboptimal. Therefore, based on rigorous empirical
 828 evaluation (see Ablation Study), we developed a hierarchical strategy designed to gracefully handle
 829 the spectrum of protein novelty:

- 830 1. **Prioritizing High-Confidence Homology:** Our analysis revealed that for generating a list of
 831 candidate GO terms, the single most reliable source is the annotation of the top homolog found
 832 by BLAST. This strategy maximizes precision while maintaining high recall (see Table 3).
- 833 2. **Integrating Domain Information:** Pfam motifs identified by InterProScan are added as a
 834 separate, complementary source of evidence, providing structural and functional context.
- 835 3. **Semantic Evidence:** Our experiments showed that ProTrek’s semantic hits, while powerful,
 836 could introduce noise when combined with high-quality GO/Pfam data. Therefore, ProTrek’s
 837 output is used as a fallback—it is only added to the context when primary sources like GO and
 838 Pfam are sparse or absent.

839 This empirically-driven, hierarchical process culminates in a final textual context engineered to be
 840 factually dense and noise-free as possible, ready for the final inference stage.

841 **A.3 CONTEXT-BASED INFERENCE**

842 The final stage transforms the structured biological evidence into a query that the LLM can process.
 843 The constructed context and the original user question are formatted into a unified prompt using a
 844 predefined template as shown in Figure 4. The LLM’s role is to act as a knowledge synthesizer.
 845 It processes the prompt, which contains a series of factual statements derived from the context.
 846 This final step leverages the LLM’s core strength in natural language understanding and reasoning,
 847 entirely bypassing the need for it to interpret the complex, low-level syntax of the raw biomolecular
 848 sequence itself.

849 **B DATASET DETAILS**

850 **B.1 PROTEIN DATASET**

851 To ensure a comprehensive and fair evaluation of our model against Evolla, we employed a multi-
 852 faceted dataset strategy. This approach incorporates not only the benchmark datasets used in the
 853 original Evolla study but also a dataset we have meticulously reconstructed to address specific limi-
 854 tations in their evaluation methodology. Our assessment is primarily based on the following datasets:

- 855 1. **Original Evolla Evaluation Dataset**

We initiated our evaluation on the original Evolla Question–Answering (QA) dataset to establish a performance baseline (Figure 4). Following the protocol from the original study (Zhou et al., 2025), this dataset’s test set is partitioned into three subsets based on sequence identity to the training set, designed to assess model generalization at varying levels of difficulty:

- **Easy:** Proteins with >30% sequence identity to the major representative clusters in the training set (covering 50% of sequences).
- **Medium:** Proteins with >30% sequence identity to the remaining training clusters.
- **Hard:** Proteins with <30% sequence identity to any training cluster.

However, despite this structured difficulty, we identified inherent limitations within the dataset. A significant portion of the questions are open-ended (e.g., “Can you share furthermore important details regarding this protein?”), for which there is no single, definitive correct answer. This ambiguity complicates evaluation, as a model’s response may be valid even if it does not align perfectly with the provided ground truth (GT). Furthermore, the GT answers were themselves generated by feeding protein database entries into a Large Language Model (LLM), rather than using the original source text, a process that may introduce stylistic biases or factual inaccuracies.

2. Our Benchmark Dataset

To overcome these evaluation challenges, we reconstructed the “Hard” subset of the Evolla dataset to create a more rigorous and objective benchmark. This subset, containing proteins with less than 30% sequence identity to the training set, is critical for assessing a model’s generalization capabilities on novel proteins. Our reconstruction was guided by the following principles:

- **Standardized Questions:** We replaced open-ended queries with a fixed set of three targeted question templates: “What is the function of this protein?”, “What is the pathway of this protein?”, and “What is the subcellular location of this protein?”
- **Conditional Question Generation:** To ensure every question has a verifiable answer, we only generated a specific question if the corresponding field (“Function,” “Pathway,” or “Subcellular location”) was explicitly present in the protein’s source database entry.
- **Authentic Ground Truth:** Crucially, our ground truth answers are direct excerpts from the protein’s database entry. Unlike the original dataset, we did not use an LLM to generate answers, thereby ensuring the objectivity and factual accuracy of the GT and creating a more reliable scoring standard.

The results for our benchmark dataset are presented in Table 1, where we compare the performance of different methods using sequence-only, context-only, and sequence+context inputs.

3. Enzyme Commission (EC) Number Dataset

To evaluate the model’s performance on a specific, structured bioinformatics task, we also utilized the EC Number dataset from the Evolla study. This task requires the model to accurately predict the functional class of enzymes, which is a standardized and important functional annotation task. Testing on this dataset allows us to gauge the model’s capabilities in handling classification-oriented protein function prediction, and is presented in Figure 9.

4. Time-Split Dataset

To investigate the model’s performance over time and the impact of sequence novelty, we curated a dataset by randomly sampling about 100 proteins for each year from 1995 to 2024 based on the first publication year. This allows us to analyze the relationship between a protein’s first publication year and the model’s LLM-Score, as shown in the “Degrading Phenomenon Across Time” section (Figure 5). The time-split dataset is valuable for understanding how well the model generalizes to older versus more recent proteins, and whether its performance degrades as the data becomes more novel, less represented in training, or more temporally distant from the model’s training data cutoff.

5. Mol-Instruction Dataset

To further evaluate the model’s ability to handle diverse protein tasks, we tested it on the Mol-Instruction benchmark (Fang et al., 2024). This benchmark provides datasets for assessing molecular understanding across a range of tasks. We specifically evaluated the model on three functionally distinct protein subsets: Catalytic Activity, General Function, and Protein Function. The performance on these datasets, as illustrated in the “Performance on Protein Function

918 Prediction” section (Figure 11), highlights the model’s ability to accurately predict protein
 919 function even across different protein families, demonstrating the robustness and flexibility of
 920 our context-driven approach.
 921

922 B.2 DNA DATASET 923

924 To assess whether our context-driven paradigm extends beyond proteomics, we evaluated its performance
 925 on a DNA-based mechanistic reasoning task. We utilized the KEGG Disease Pathway dataset
 926 curated by BioReason (Fallahpour et al., 2025), which provides a unique benchmark for connecting
 927 genomic variants to disease phenotypes through multi-step biological pathways. Each entry in the
 928 dataset consists of a reference and a variant DNA sequence, the associated KEGG pathway definition,
 929 and the ground-truth disease outcome. The task requires the model to reason from the mutation
 930 and its functional context to predict the correct disease.

931 We designed three experimental setups to investigate the impact of different data configurations
 932 on the model’s performance. The first setup included only pathway-related contextual information
 933 (context-only). The second setup incorporated both the pathway context and the raw DNA sequence
 934 data (context and sequence). The third setup focused solely on the DNA sequence itself (sequence-
 935 only). These configurations allowed us to evaluate the effect of using context, sequence, or both on
 936 the model’s ability to predict mutations.

937 The KEGG dataset’s comprehensive pathway data, paired with precise mutation annotations, pro-
 938 vided a solid foundation for designing these experiments. By varying the inclusion of sequence and
 939 context information, we aimed to assess the model’s ability to predict the effects of DNA mutations
 940 based on both pathway context and raw sequence data.

941 C EVALUATION METRIC 942

943 To conduct a comprehensive and multi-dimensional assessment of our model’s performance, we
 944 designed specific evaluation metrics tailored to each of our distinct tasks.
 945

946 C.1 LLM-SCORE FOR GENERAL PROTEIN QA TASKS 947

948 For the open-ended protein question–answering task, traditional metrics based on lexical overlap
 949 (e.g., BLEU, ROUGE) are inadequate for assessing the semantic accuracy and factual consistency
 950 of generated answers. To address this, we adopted an automated evaluation methodology leveraging
 951 a LLM as an adjudicator, which we term the LLM-Score. The core principle of this metric is to use
 952 a powerful, independent third-party LLM (in this case, a DeepSeek-V3 (Liu et al., 2024a) model) to
 953 score the quality of our model’s generated answer against the ground truth. The evaluation process
 954 is as follows:
 955

- 956 1. **Prompt Construction:** We embed the generated answer and the ground truth answer into a
 957 carefully designed prompt template. This prompt instructs the adjudicator LLM to act as an
 958 expert biologist and perform a holistic evaluation based on factual accuracy. The exact prompt
 959 is shown below.
 960

961 LLM-Score Adjudicator Prompt 962

963 As an expert biologist, you are assigned to check one paragraph
 964 is aligned with facts or not. You will receive some facts, and
 965 one paragraph. Score the paragraph between 0 to 100.
 966 The score should be the format of {"score": score}

967 -----
 968 **Here's the facts:**
 969 [Ground Truth Text from Database]

970 -----
 971 **Here's the paragraph:**
 972 [Generated Answer from Model to be Scored]

- 972 2. **Score Generation:** The adjudicator LLM processes the prompt and returns a numerical score
 973 on a scale from 0 to 100, where a higher score indicates that the generated answer is of higher
 974 quality and more closely aligned with the ground truth.
 975 3. **Score Extraction and Aggregation:** A robust parsing function extracts the numerical score
 976 from the LLM’s textual response. The model’s final performance on the dataset is reported as
 977 the *average LLM-Score* across all test samples.
 978

979 This approach moves beyond surface-level text matching to provide a deeper, more semantically
 980 aware assessment of the model’s ability to comprehend and articulate biological knowledge.
 981

982 C.2 HIERARCHICAL METRICS FOR EC NUMBER PREDICTION

983 The Enzyme Commission (EC) number is a four-level hierarchical classification system (e.g.,
 984 A.B.C.D). A proficient model should be rewarded not only for predicting the exact four-digit code
 985 but also for correctly identifying the broader functional classes at higher levels of the hierarchy.
 986 Therefore, an exact-match accuracy metric at a single level is insufficient.
 987

988 To capture this, we implemented a more nuanced, hierarchical evaluation scheme. We calculate
 989 **F1-Score** at each of the four functional levels.

990 The methodology is as follows:
 991

- 992 • **Hierarchical Matching:** To evaluate performance at `level-N`, all predicted and ground truth
 993 EC numbers are truncated to their first `N` digits for comparison. For example, at `level-3`, a
 994 prediction of 1.2.3.5 is considered a match for a ground truth of 1.2.3.4.
- 995 • **Multi-Label Formulation:** As a single protein can be associated with multiple EC numbers,
 996 the task is treated as a multi-label classification problem.
- 997 • **Micro-Averaging:** We compute the total number of True Positives (TP), False Positives (FP),
 998 and False Negatives (FN) by aggregating their counts over the entire test set. Global Precision,
 999 Recall, and F1-Score are then calculated from these aggregate sums.

$$1000 \quad 1001 \quad \text{Precision}_{\text{micro}} = \frac{\sum \text{TP}_i}{\sum \text{TP}_i + \sum \text{FP}_i} \quad (7)$$

$$1002 \quad \text{Recall}_{\text{micro}} = \frac{\sum \text{TP}_i}{\sum \text{TP}_i + \sum \text{FN}_i} \quad (8)$$

$$1003 \quad \text{F1}_{\text{micro}} = 2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \quad (9)$$

1004 This suite of metrics provides a comprehensive view of the model’s predictive accuracy at varying
 1005 degrees of granularity and effectively handles the multi-label nature of the data, offering a more
 1006 equitable measure of true performance.
 1007

1008 D QUANTITATIVE BENCHMARK: EC NUMBER PREDICTION

1009 To further validate our central thesis on the primacy of context over sequence, we introduce a quan-
 1010 titative benchmark: EC number prediction. This task provides an objective, verifiable measure of
 1011 a model’s ability to understand a protein’s precise biochemical function. The hierarchical nature of
 1012 EC numbers allows us to evaluate performance at four increasing levels of specificity (from 1-digit
 1013 to 4-digit matches), with the F1-Score serving as our primary metric.
 1014

1015 We compare two categories of models: “Sequence-Only” models, which include both general-
 1016 purpose LLMs and CLEAN (Yu et al., 2023) (a model specifically trained for this task), and
 1017 “Context-Driven” models, which leverage the contextual information as described in the main text.
 1018 The comparative performance is visualized in Figure 9.
 1019

1020 The results presented in Figure 9 are unequivocal and offer several key insights:
 1021

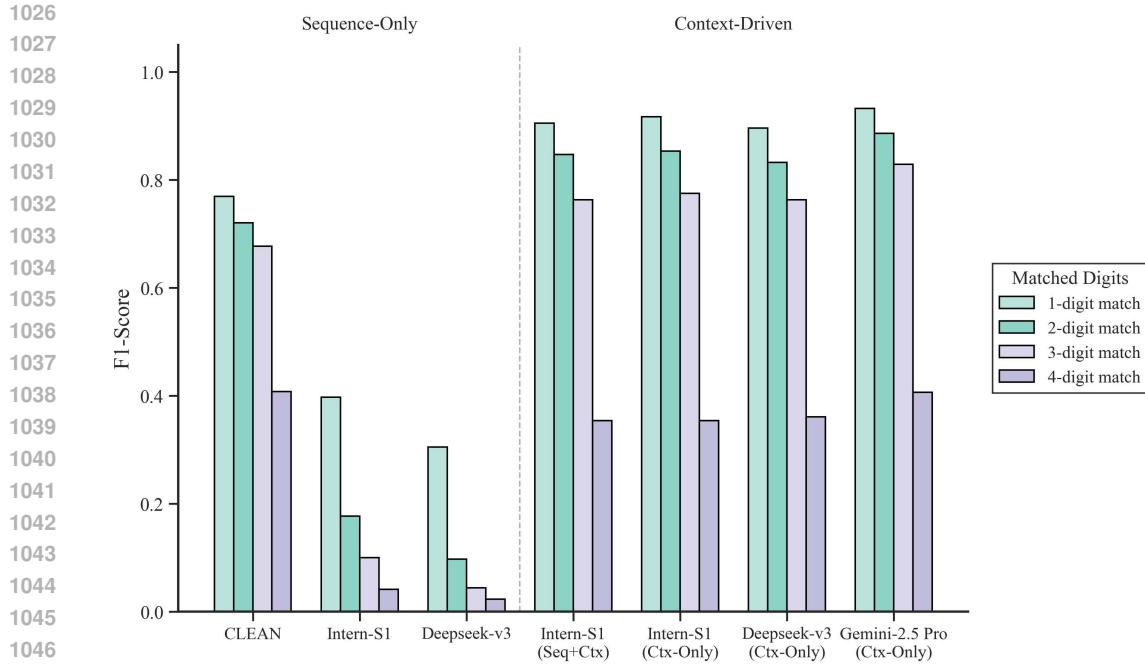


Figure 9: **Performance on EC Number Prediction (F1-Score).** The plot is divided into Sequence-Only models (left) and Context-Driven models (right). A clear and dramatic performance gap is visible between the two groups. Context-driven approaches significantly outperform even specialized sequence-based models like CLEAN, especially at higher levels (3 and 4 digits).

- **Failure of Sequence-Only LLMs:** In the Sequence-Only setting, both general-purpose LLM DeepSeek-V3 and Sci-LLM Intern-S1 perform poorly. Their F1-scores plummet as the required specificity increases, demonstrating their inability to decipher complex enzymatic function from raw sequence data alone. This reinforces our “lost in tokenization” hypothesis.
- **Context Outperforms Specialization:** CLEAN, a model specifically trained on sequences for EC prediction, establishes a respectable baseline. However, every model in the Context-Driven category outperforms CLEAN across the first three levels of precision (1-digit, 2-digit, and 3-digit). This demonstrates that providing high-level context to a general model is more effective than training a specialized model on sequence data for these levels.
- **Robustness of the Context-Driven Approach:** While all models show a natural decline in performance as the task becomes harder (from 1-digit to 4-digit prediction), the context-driven models exhibit a much more graceful degradation. Gemini-2.5 Pro, using only context, achieves the highest F1-score of 0.406 on the most challenging 4-digit prediction task, a score comparable to the specialized CLEAN model’s performance (0.408).

In summary, this quantitative benchmark provides strong, direct evidence that high-level biological context is a far more effective and reliable representation of protein function for LLMs than the raw amino acid sequence. It enables general-purpose models to excel at highly specific bioinformatics tasks without needing task-specific architectures or fine-tuning.

E ABLATION STUDY

To dissect the contribution of each component within our framework, we conducted a comprehensive ablation study. The experiments were performed on our benchmark dataset, which is ideal for evaluating generalization, as it contains proteins with less than 40% sequence identity to the Evolla training set. We systematically evaluated the performance by providing different combinations of contextual information—Pfam, GO, and ProTrek—to the DeepSeek model. The results, summarized in Table 3, reveal the individual and synergistic effects of these components.

1080
 1081 Table 3: Ablation study of context components on our benchmark dataset. Scores reflect the model’s
 1082 performance when provided with different combinations of contextual information. Our final, con-
 1083 ditional approach yields the best result.

1084	Context Components Provided	LLM Score
Single Components		
1086	Pfam only	74.90
1087	GO only	84.02
1088	ProTrek only	66.44
Pairwise Combinations		
1090	Pfam + GO	84.60
1091	Pfam + ProTrek	77.00
1092	GO + ProTrek	77.78
Full Combinations		
1094	Pfam + GO + ProTrek (Unconditional)	81.56
1095	Pfam + GO + ProTrek (Conditional)	84.99

1097
 1098 **Analysis of Individual Components** The results from single-component experiments establish a
 1099 clear hierarchy of information value. Gene Ontology (GO) annotations emerge as the most powerful
 1100 single source of context, achieving a high score of 84.02 on its own. Pfam provides a moderately
 1101 strong signal, scoring 74.90. In contrast, ProTrek alone is the least informative component, with a
 1102 score of 66.44, suggesting its raw output may be noisy or less directly useful for functional queries.

1103
 1104 **Synergistic and Antagonistic Effects in Combinations** Combining Pfam and GO yields a score
 1105 of 84.60, a slight improvement over GO alone, indicating a positive, synergistic relationship where
 1106 Pfam provides complementary information. However, a critical observation arises when combining
 1107 components with ProTrek. Both ‘Pfam + ProTrek’ (77.00) and ‘GO + ProTrek’ (77.78) perform
 1108 worse than their stronger counterparts (Pfam and GO, respectively) alone. This trend is amplified
 1109 when all three are combined unconditionally (‘Pfam + GO + ProTrek’), resulting in a score of 81.56,
 1110 which is substantially lower than ‘Pfam + GO’. This strongly suggests that naively adding ProTrek’s
 1111 information introduces noise that dilutes the high-quality signals from Pfam and GO, ultimately
 1112 degrading the model’s performance.

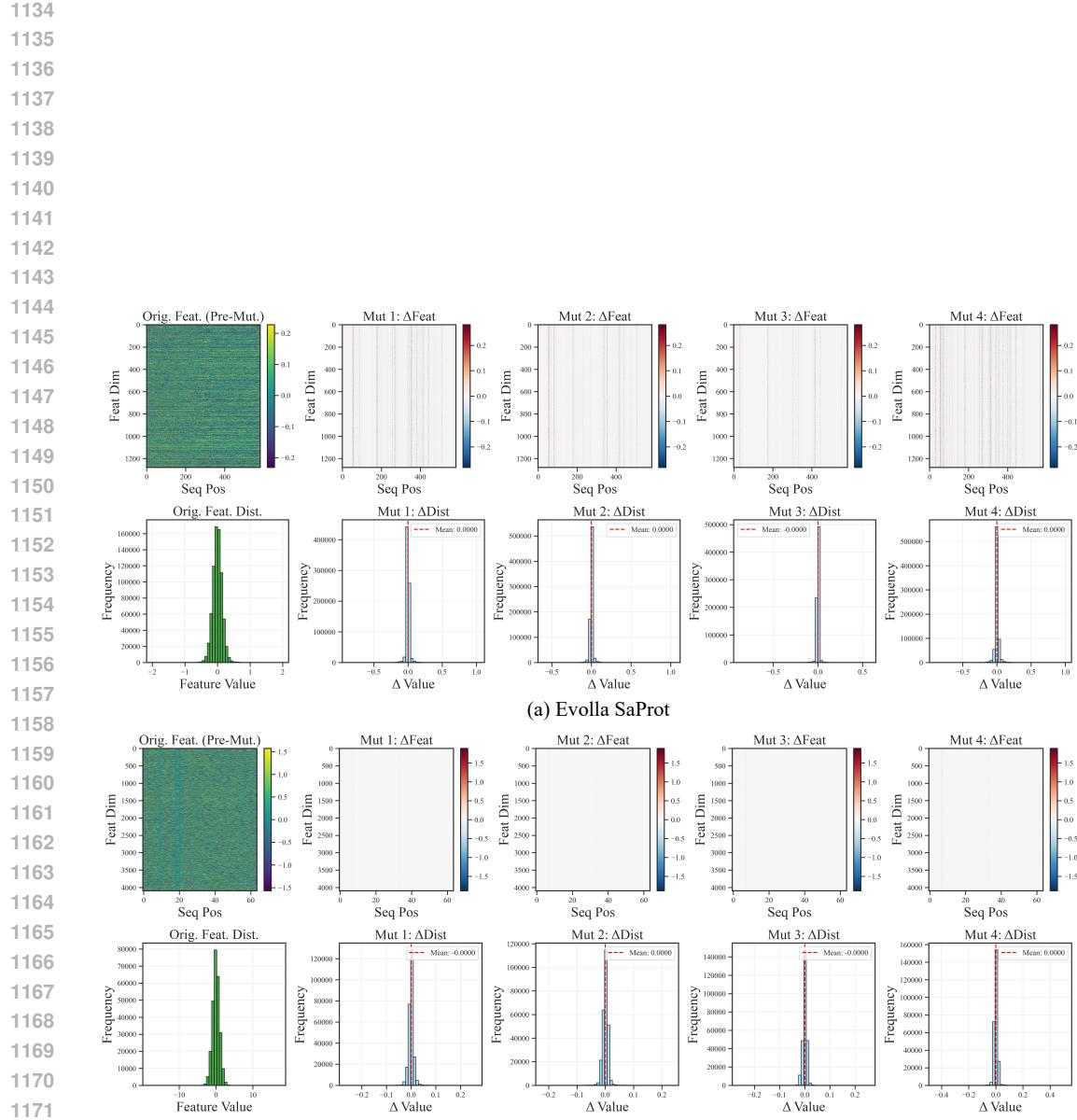
1113
 1114 **Justification for the Conditional Strategy** Based on this insight, we implemented our final, con-
 1115 ditional strategy: ProTrek information is only included as a fallback when both Pfam and GO annota-
 1116 tions are unavailable for a given protein. This intelligent inclusion prevents ProTrek from interfer-
 1117 ing with higher-quality data while still providing a baseline of information for sparsely annotated pro-
 1118 teins. As shown in the final row of Table 3, this conditional approach achieves the highest score
 1119 of 84.99. It effectively captures the synergy of Pfam and GO while mitigating the negative, noisy
 1120 impact of ProTrek, thus justifying its selection as our final methodology.

1121 F IMPACT OF SEMANTIC ALIGNMENT ON MUTATION SENSITIVITY

1122
 1123 A critical capability for any protein model is the ability to detect and represent the effects of small se-
 1124 quence variations, such as point mutations. To investigate how the internal mechanisms of sequence-
 1125 as-modality models like Evolla handle such changes, we conducted an analysis on the feature repre-
 1126 sentations before and after its Q-Former alignment module.

1127
 1128 We introduced a series of mutations (from 1 to 4 differing sites) into a sample protein sequence.
 1129 We then extracted the resulting feature embeddings at two key stages: (1) directly from the SaProt
 1130 protein encoder, and (2) after they had been processed by the Q-Former. The difference between
 1131 the pre-mutation (wild-type) and post-mutation embeddings at each stage was then visualized and
 1132 quantified. The results are presented in Figure 10.

1133 **High Sensitivity at the Protein Encoder Stage** As expected, the SaProt protein encoder is highly
 1134 sensitive to sequence mutations. The visualizations (Figure 10, top row) show clear, localized



1174 Figure 10: Effect of mutations on internal representations of Evolla. The top row shows feature
1175 differences from the **SaProt encoder**, and the bottom row from the **Q-Former**. Heatmaps visualize
1176 the difference vector (‘mutated - original’). SaProt’s representation is clearly perturbed by
1177 mutations, showing localized and significant changes. In contrast, the Q-Former’s output shows almost
1178 no change, indicating that the alignment process erases the fine-grained signal of the mutation.

1179
1180
1181
1182
1183
1184
1185
1186
1187

1188 changes in the feature map corresponding to the mutation sites. Quantitatively, while the cosine
 1189 similarities between pre- and post-mutation embeddings remain high (0.980–0.995), the Euclidean
 1190 distances are substantial (ranging from 13.3 to 26.2). This confirms that the encoder accurately cap-
 1191 tures the perturbation caused by the mutation, altering the feature vector in a meaningful way. This
 1192 sensitivity is the foundation required for any downstream mutation effect analysis.
 1193

1194 **Loss of Sensitivity After Q-Former Alignment** A starkly different picture emerges after the
 1195 features pass through the Q-Former. The difference heatmaps (Figure 10, bottom row) are almost
 1196 entirely uniform, indicating a negligible change between the wild-type and mutated representations.
 1197 This visual observation is confirmed by the quantitative metrics: the cosine similarities are nearly
 1198 perfect (approaching 1.0, e.g., > 0.9999), and the Euclidean distances (5.8–9.9) are significantly
 1199 smaller than those observed from the encoder.
 1200

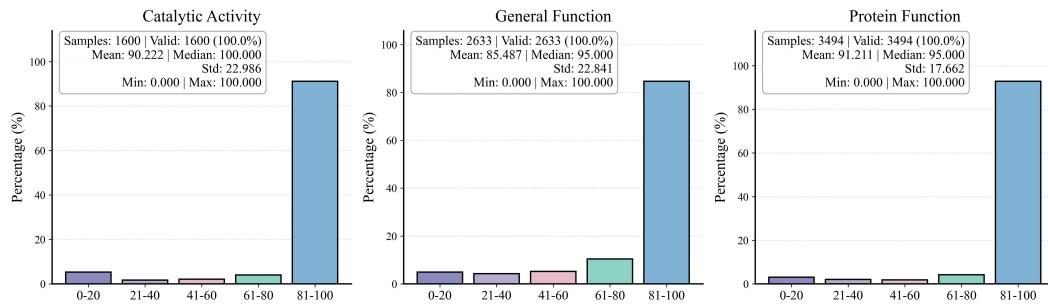
1201 **Implications for Downstream Tasks** This analysis reveals a critical limitation of the sequence-as-
 1202 modality paradigm employed by models like Evolla. The Q-Former, in its role of compressing and
 1203 aligning the detailed protein features into a fixed set of tokens for the language model, effectively
 1204 “smooths out” or discards the fine-grained information related to single point mutations. While this
 1205 may be sufficient for generating high-level functional descriptions, it renders the final representation
 1206 insensitive to the subtle yet critical differences that underpin tasks like mutation effect prediction,
 1207 disease variant analysis, and protein engineering. This inherent loss of information at the alignment
 1208 stage explains why such architectures are fundamentally ill-suited for these precision tasks.
 1209

1210 G GENERALIZABILITY ACROSS BIOMOLECULAR TYPES

1211 To demonstrate the broad applicability and robustness of our context-driven methodology, we eval-
 1212 uated its performance on standard benchmarks beyond our primary QA dataset. This tests the ap-
 1213 proach on different tasks and different biomolecular types.
 1214

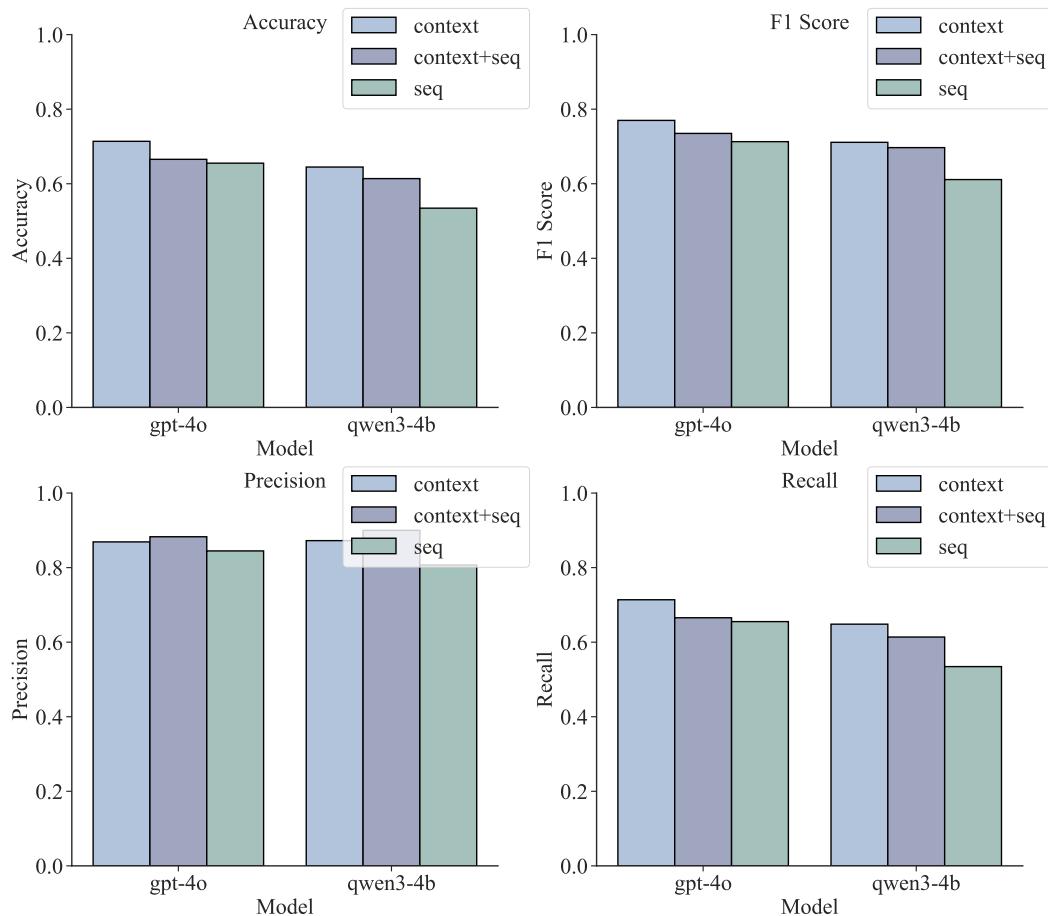
1215 G.1 PERFORMANCE ON PROTEIN FUNCTION PREDICATION

1216 We first evaluate our approach on the protein classification tasks from the Mol-Instruction bench-
 1217 mark (Fang et al., 2024). This benchmark contains curated datasets for assessing molecular under-
 1218 standing. We specifically tested on three functionally distinct protein subsets: Catalytic Activity,
 1219 General Function, and Protein Function. The performance is shown in Figure 11.
 1220



1232 Figure 11: Performance on Mol-Instruction Protein Classification. Score distributions of our
 1233 context-driven method on three sub-datasets. The results show consistently high performance across
 1234 all categories, with mean scores above 85 and the vast majority of answers ($> 84\%$) falling into the
 1235 highest score bracket (81 – 100), demonstrating robust generalization.
 1236

1237 As illustrated in Figure 11, our method achieves excellent performance across all three predica-
 1238 tion tasks. The mean score were exceptionally high: 91.2 for Protein Function, 90.2 for Catalytic
 1239 Activity, and 85.5 for General Function. The score distributions are heavily skewed towards the
 1240 maximum, with over 84% of answers in all three tasks receiving a score in the 81-100 range. This
 1241 demonstrates that our method is not only effective on our QA benchmark but also generalizes ro-
 bustly to standard, multi-category protein function predication tasks, validating its broad utility.
 1242

1242 G.2 PERFORMANCE ON DNA MUTATION PREDICTION
1243

1274 Figure 12: Model performance on the KEGG DNA dataset under different input configurations.
1275 Both Qwen3-4B and GPT-4o were evaluated across context-only, sequence+context, and combined
1276 inputs, with both models showing consistent improvements when using context-based inputs
1277 compared to sequence-based approaches.

1278 We evaluated two powerful generalist LLMs, GPT-4o and Qwen3-4B, across our three standard
1279 input configurations: Context-Only, Sequence + Context, and Sequence-Only. To provide a com-
1280 prehensive view of performance, we measured not only classification accuracy but also F1 Score,
1281 Precision, and Recall, accounting for potential class imbalances in the dataset. The results, pre-
1282 sented in Figure 12, strongly corroborate our findings from the protein-based tasks and confirm the
1283 paradigm’s generalizability. For both models, the Context-Only configuration consistently achieved
1284 the highest scores across all four evaluation metrics. Crucially, the “informational noise” effect of
1285 raw sequences persists in the DNA domain. The Sequence + Context configuration consistently un-
1286 derperformed the Context-Only setup, indicating that the models were again “lost in tokenization,”
1287 struggling to integrate the low-level signal from the raw DNA sequence with the clear, high-level
1288 context. The Sequence-Only approach yielded the poorest results, confirming that atomic tokeniza-
1289 tion of nucleotide sequences is insufficient for complex biological reasoning.

1291 H INDEPENDENCE FROM CLUSTERING METRICS
1292

1293 To verify that our conclusions are robust and not contingent on a single definition of protein families,
1294 we evaluated all embeddings against two distinct ground-truth labeling schemes: UniClust50 and
1295 UniClust30. These standards group proteins at 50% and a stricter 30% sequence identity threshold,

1296 respectively, providing different granularities for functional classification. While both are generated
 1297 using the MMseqs2 algorithm, they represent conceptually different criteria for defining protein
 1298 homology. We performed the same hierarchical clustering analysis for each representation against
 1299 both sets of labels. The resulting ARI scores are presented in Table 4.

1302 **Table 4: Performance (ARI) against ground-truth labels from UniClust50 and UniClust30.**

1303 Model / Representation Stage	1304 ARI (vs. UniClust50)	1305 ARI (vs. UniClust30)
1306 Ours	0.958	0.958
1307 Evolla Encoder (SaProt)	0.945	0.945
1308 Evolla Alignment (Q-Former)	0.916	0.916
1309 Evolla Decoder (Final)	0.809	0.809
1310 Intern-S1 8B	0.690	0.690
NatureLM	0.492	0.492

1312 As shown in Table 4, the performance hierarchy and the specific ARI scores remain identical across
 1313 both labeling schemes. While the identical results suggest that our test set’s structure is stable
 1314 across these two identity thresholds, the key insight is the unwavering consistency of our central
 1315 observations. Specifically, three conclusions hold firm regardless of the ground-truth definition:

- 1317 • Our context-driven approach consistently achieves the highest functional separation (ARI
 1318 0.958).
- 1319 • The ‘semantic misalignment’ within Evolla is consistently demonstrated by the progressive
 1320 drop in ARI from 0.945 to 0.809.
- 1321 • The ‘weak representation’ of sequence-to-language models (Intern-S1 and NatureLM) is con-
 1322 sistent with their significantly lower scores.

1325 This stability across different conceptual standards for protein families validates our findings, con-
 1326 firming that the tokenization dilemma and the superiority of our context-driven solution are funda-
 1327 mental phenomena, not artifacts of a particular evaluation metric.

1329 I QA EXAMPLES

1332 To provide concrete illustrations of the performance differences reported in the main paper, this
 1333 section presents qualitative examples from our benchmark. We analyze the outputs for a single
 1334 protein (UniProt ID: A6LHQ9) across the three input modalities: Context-Only, Sequence-Only,
 1335 and Sequence + Context. These case studies offer tangible insights into why our context-driven
 1336 approach excels and how raw sequence inputs can be detrimental.

1338 I.1 COMPARATIVE ANALYSIS ON PROTEIN A6LHQ9

1340 The following examples use the same protein and question (“What is the function of this protein?”)
 1341 to demonstrate how the input modality directly impacts the model’s reasoning and final output. The
 1342 ground truth is: “Putative component of the fimbrium tip. Fimbriae are filamentous appendages on
 1343 the cell surface that mediate cell adhesion and biofilm formation”.

1344 I.1.1 CASE 1: CONTEXT-ONLY (SUCCESS)

1346 In the Context-Only setting, the model receives curated information from Pfam. The model demon-
 1347 strates a strong ability to synthesize this structured knowledge into a comprehensive and accurate
 1348 answer, achieving a perfect score.

1350

Protein A6LHQ9: Context-Only

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

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

Question:

What is the function of this protein?

context Provided:

Conserved Domains (from Pfam):

- **Pfam PF06321:** This family consists of several *Porphyromonas gingivalis* major fimbrial subunit protein (FimA) sequences. Fimbriae of *Porphyromonas gingivalis*, a periodontopathogen, play an important role in its adhesion to and invasion of host cells. The fimA genes encoding fimbillin (FimA), a subunit protein of fimbriae, have been classified into five types, types I to V, based on nucleotide sequences. It has been found that type II FimA can bind to epithelial cells most efficiently through specific host receptors [[cite:PUB00010404]]. Human dental plaque is a multispecies microbial biofilm that is associated with two common oral diseases, dental caries and periodontal disease. There is an inter-species contact-dependent communication system between *P. gingivalis* and *S. cristatus* that involves the Arc-A enzyme [[cite:PUB00069820]].
- **Pfam PF22449:** This domain is found at the C-terminal end of the putative fimbrium tip subunit Fim1F from *Parabacteroides distasonis*, which contain an N-terminal domain ([pfam:PF06321]) and a slightly larger C-terminal domain (this entry) with a transthyretin-like fold that contains seven core beta-strands arranged in two beta-sheets and an extra conserved 'appendage' of two amphipathic beta-strands [[cite:PUB00080711]].

Answer:

The protein in question, which incorporates the domains PF06321 (fimbrial subunit FimA) and PF22449 (C-terminal transthyretin-like domain), functions as a **structural component of bacterial fimbriae** with a primary role in **adhesion to host cells and surfaces**. Here's a detailed breakdown:

- **PF06321 (FimA-like domain):** Found in fimbrial subunits of *Porphyromonas gingivalis* (a periodontal pathogen), this domain mediates host cell adhesion and invasion. - Type II FimA (a variant within this family) binds efficiently to epithelial cells via specific host receptors, highlighting its role in pathogenicity and biofilm formation (e.g., in dental plaque).
- **PF22449 (transthyretin-like domain):** Present at the C-terminus of fimbrial tip subunits (e.g., Fim1F in *Parabacteroides distasonis*), this domain stabilizes the fimbrial structure through its transthyretin-like fold (seven beta-strands forming two beta-sheets, plus an amphipathic 'appendage'). - The transthyretin-like fold may enhance structural integrity or enable additional interactions, such as binding to host molecules or other bacterial proteins.

Overall Function: The protein acts as a fimbrial structural subunit, critical for assembling adhesive fimbriae. These appendages facilitate bacterial attachment to host tissues (e.g., oral epithelial cells) and surfaces, enabling colonization, biofilm formation, and inter-species interactions (e.g., with *Streptococcus cristatus* via the Arc-A system). The combination of PF06321 and PF22449 domains likely optimizes both adhesion efficiency and structural stability, contributing to the

1404
 1405 pathogenicity of periodontal bacteria. Key References: -
 1406 PF06321: Binding to host receptors (PUB00010404). - PF22449:
 1407 Structural role in fimbriae (PUB00080711). - Inter-species
 1408 communication (PUB00069820). -----
 1409 **Ground Truth:**
 1410 Putative component of the fimbrium tip. Fimbriae are filamentous
 1411 appendages on the cell surface that mediate cell adhesion and
 1412 biofilm formation.
 1413 -----
 1414 **Result:**
 1415 LLM-Score: 100

I.1.2 CASE 2: SEQUENCE-ONLY (TOTAL FAILURE)

1416 When provided with only the raw amino acid sequence, the model fails completely. It hallucinates
 1417 a function (Chitin-Binding Module) that is entirely incorrect, likely by misinterpreting spurious
 1418 patterns in the tokenized sequence. This is a stark example of the “lost in tokenization” phenomenon.
 1419

1420
 1421 Protein A6LHQ9: Sequence-Only
 1422
 1423 You are a senior systems biologist. Analyze the input
 1424 information to answer the given question.
 1425 -----
 1426 **Question:**
 1427 What is the function of this protein?
 1428 -----
 1429 **Sequence Provided:**
 1430 MRFNVVLFMLIVALLGGLSTCSSEVPIGFDTDELSFDMSLVLNTGDMQTKASDPNYTYATTEEL
 1431 TIQNCHVAVFDKDGKRIYFKNFYSKDLGEMKTIGNLSGYELQLEGVRTFGKEDKKVSVLVVANA
 1432 NNANNNSPFDNLTTYDGVDNSYTAKTIAKGPVTASLLVKICKSETTLKYNQDNAPVTVSLIQLSA
 1433 KIEYTGVYKKENGELLEGFSLTKVAGLNASSKITIFNTSAVENGAFFSDLAYPTTKPVTFYTYEI
 1434 SDAFKEVILSVQSGVEPKYEPFPANKFIKGNNYYRIKGLKSSTEIEWVLENVEDKEVTLDPE
 1435 -----
 1436 **Answer:**
 1437 Based on the amino acid sequence provided, this protein is **very**
 1438 **likely a Chitin-Binding Module (specifically, a CBM12 domain).**
 1439 Here's a breakdown of the evidence:
 1440 • **Signal Peptide:** The sequence starts with MRFNVVLFMLIVALLGGLSTCSS,
 1441 which is a classic signal peptide. This indicates the protein
 1442 is secreted from the cell.
 1443 • **CBM12 Signature Motifs:** The sequence contains the highly
 1444 characteristic motifs of the Carbohydrate-Binding Module family
 1445 12 (CBM12):
 1446 - DGDG (appears twice: DGDG and DGDG).
 1447 - NGAN (appears: NGAN).
 1448 - TGS (appears: TGS).
 1449 - QGN (appears: QGN).
 1450 - Multiple conserved cysteine residues (C) that likely form
 1451 stabilizing disulfide bonds.
 1452 • **Function of CBM12:**
 1453 - **Primary Function:** Carbohydrate binding, specifically to
 1454 N-acetylglucosamine (GlcNAc) residues. GlcNAc is the building
 1455 block of chitin (a major component of fungal cell walls,
 1456 insect exoskeletons, and crustacean shells) and is also found
 1457 in various glycoproteins and glycolipids (like hyaluronan).

1458
 1459 - **Role in Proteins:** CBM12 domains are typically found
 1460 attached to enzymes (like chitinases, chitobiosidases, or
 1461 N-acetylglucosaminidases) or structural proteins. Their role
 1462 is to target the catalytic domain to its specific substrate
 1463 (chitin or other GlcNAc-containing polysaccharides) and
 1464 increase the enzyme's efficiency by bringing it into close
 1465 proximity with the substrate.
 1466 - **Binding Specificity:** CBM12 modules generally bind to
 1467 short stretches of GlcNAc residues, including chitobiose
 1468 (GlcNAc-GlcNAc) and chitotriose (GlcNAc-GlcNAc-GlcNAc), and
 1469 show weaker binding to isolated GlcNAc monomers. They do not
 1470 bind cellulose.
 1471 In summary, the primary function of this protein is to act as
 1472 a Chitin-Binding Module (CBM12). Its job is to specifically
 1473 recognize and bind to N-acetylglucosamine (GlcNAc)-containing
 1474 carbohydrates, particularly chitin, facilitating the action
 1475 of associated enzymes (like chitinases) or contributing to the
 1476 structural integrity/function of proteins interacting with chitin
 1477 or related polysaccharides.
 1478 Typical Context: Proteins containing CBM12 domains are very
 1479 common in fungi, where they play crucial roles in cell wall
 1480 remodeling, nutrient acquisition (degrading chitin from other
 1481 organisms), and pathogenicity. They are also found in bacteria
 1482 and some other organisms.
 1483 -----
 1484 **Ground Truth:**
 1485 Putative component of the fimbrium tip. Fimbriae are filamentous
 1486 appendages on the cell surface that mediate cell adhesion and
 1487 biofilm formation.
 1488 -----
 1489 **Result:**
 1490 LLM-Score: 0

I.1.3 CASE 3: SEQUENCE + CONTEXT (NOISE EFFECT)

1491 In the combined setting, the model produces a correct but slightly less focused answer than the
 1492 Context-Only version. The inclusion of the sequence acts as informational noise, causing a minor
 1493 degradation in answer quality and a corresponding drop in the score. This supports our quantitative
 1494 finding that the sequence is not merely unhelpful but can be actively detrimental.

1495 Protein A6LHQ9: Sequence + Context
 1496
 1497 You are a senior systems biologist. Analyze the input
 1498 information to answer the given question.
 1499 -----
 1500 **Question:**
 1501 What is the function of this protein?
 1502 -----
 1503 **Context Provided:** [Same as Case 1]
 1504 **Sequence Provided:** [Same as Case 2]
 1505 -----
 1506 **Answer:**
 1507 The protein in question, based on the sequence and associated
 1508 motifs (PF06321 and PF22449), functions as a **fimbrial subunit**
 1509 **protein** involved in bacterial adhesion and host-microbe
 1510 interactions. Here's a detailed breakdown of its role:
 1511 • **Fimbrial Structure and Assembly:** -The PF06321 domain
 1512 (N-terminal) corresponds to the major fimbrial subunit FimA

1512

of *Porphyromonas gingivalis*. This domain is critical for forming the structural backbone of fimbriae, which are hair-like appendages on the bacterial surface. - The PF22449 domain (C-terminal) contains a transthyretin-like fold with a conserved ‘‘appendage’’ of amphipathic beta-strands. This domain likely stabilizes the fimbrial tip structure and may mediate specific interactions with host cells or other microbial species.

1519

- **Adhesion and Invasion:** - Fimbriae in *P. gingivalis* (a periodontal pathogen) are essential for host cell adhesion and invasion. The type II FimA variant (relevant to PF06321) binds efficiently to epithelial cells via specific host receptors, facilitating colonization and immune evasion. - The transthyretin-like fold in PF22449 may enhance binding specificity or structural resilience, enabling the fimbriae to withstand mechanical stress in the oral environment.

1526

- **Biofilm Formation and Interspecies Communication:** - Fimbriae contribute to the formation of multispecies biofilms (e.g., dental plaque), which are linked to oral diseases like periodontitis. The protein may mediate interbacterial adhesion, as seen in the interaction between *P. gingivalis* and *Streptococcus cristatus* via the Arc-A enzyme system. - The amphipathic beta-strand ‘‘appendage’’ in PF22449 could play a role in direct contact-dependent signaling or physical interactions with other microbes. item **Pathogenicity:** - By enabling adhesion to host tissues and other bacteria, this protein promotes the establishment of polymicrobial infections, contributing to the progression of periodontal disease and potentially systemic inflammation.

1538

Conclusion: The protein is a fimbrial subunit that assembles into adhesive structures critical for bacterial attachment to host cells, biofilm formation, and interspecies interactions in the oral microbiome. Its dual-domain architecture (PF06321 + PF22449) ensures both structural integrity and functional specificity in these processes.

1543

Ground Truth:

Putative component of the fimbrium tip. Fimbriae are filamentous appendages on the cell surface that mediate cell adhesion and biofilm formation.

1547

Result:

LLM-Score: **95**

1550

1551

1552

1553 The comparative analysis of protein A6LHQ9 clearly illustrates the core tenets of our findings: context is the primary driver of accurate biological reasoning, while raw sequence data can lead to 1554 catastrophic failures or introduce performance-degrading noise.

1556

1557

1558

1559

1560 I.2 SHOWCASE OF OUR METHOD’S SUCCESS CASES

1561

1562

1563 Having established the superiority of the context-driven paradigm, we now showcase its robustness 1564 and versatility across the three primary query types in our benchmark: molecular function, metabolic 1565 pathway, and subcellular localization. The following examples demonstrate the model’s ability to consistently generate high-quality, accurate, and detailed answers for each category.

1566 I.2.1 EXAMPLE 1: FUNCTION PREDICTION
15671568 The example of PDB ID: P0DJ91
15691570 You are a senior systems biologist. Analyze the input
1571 information to answer the given question.
15721573 **Question:**
1574 What is the function of this protein?
15751576 **Conserved Domains (from Pfam):**
1577

- 1578 • **PF03549:** Intimin and its translocated intimin receptor
(Tir) are bacterial proteins that mediate adhesion between
mammalian cells and attaching and effacing (A/E) pathogens.
1579 A unique and essential feature of A/E bacterial pathogens is
1580 the formation of actin-rich pedestals beneath the intimately
1581 adherent bacteria and localised destruction of the intestinal
1582 brush border. The bacterial outer membrane adhesin, intimin,
1583 is necessary for the production of the A/E lesion and diarrhoea.
1584 The A/E bacteria translocate their own receptor for intimin,
1585 Tir, into the membrane of mammalian cells using the type III
1586 secretion system. The translocated Tir triggers additional host
1587 signalling events and actin nucleation, which are essential for
1588 lesion formation [[cite:PUB00006623]]. This family represents
1589 the Tir intimin-binding domain (Tir IBD) which is needed to
1590 bind intimin and support the predicted topology for Tir, with
1591 both N- and C-terminal regions in the mammalian cell cytosol
1592 [[cite:PUB00014294]].

- 1593 • **PF07489:** Intimin and its translocated intimin receptor (Tir)
1594 are bacterial proteins that mediate adhesion between mammalian
1595 cells and attaching and effacing (A/E) pathogens. A unique and
1596 essential feature of A/E bacterial pathogens is the formation
1597 of actin-rich pedestals beneath the intimately adherent bacteria
1598 and localised destruction of the intestinal brush border. The
1599 bacterial outer membrane adhesin, intimin, is necessary for the
1600 production of the A/E lesion and diarrhoea. The A/E bacteria
1601 translocate their own receptor for intimin, Tir, into the
1602 membrane of mammalian cells using the type III secretion system.
1603 The translocated Tir triggers additional host signalling
1604 events and actin nucleation, which are essential for lesion
1605 formation [[cite:PUB00006623]]. This family represents the Tir
1606 C-terminal domain which has been reported to bind uninfected
1607 host cells and beta-1 integrins although the role of intimin
1608 binding to integrins is unclear. This intimin C-terminal
1609 domain has also been shown to be sufficient for Tir recognition
1610 [[cite:PUB00014294]].

- 1611 • **PF07490:** Intimin and its translocated intimin receptor (Tir)
1612 are bacterial proteins that mediate adhesion between mammalian
1613 cells and attaching and effacing (A/E) pathogens. A unique and
1614 essential feature of A/E bacterial pathogens is the formation
1615 of actin-rich pedestals beneath the intimately adherent bacteria
1616 and localised destruction of the intestinal brush border. The
1617 bacterial outer membrane adhesin, intimin, is necessary for the
1618 production of the A/E lesion and diarrhoea. The A/E bacteria
1619 translocate their own receptor for intimin, Tir, into the
membrane of mammalian cells using the type III secretion system.
The translocated Tir triggers additional host signalling events
and actin nucleation, which are essential for lesion formation
[[cite:PUB00006623]]. This family represents the Tir N-terminal
domain which is involved in Tir stability and Tir secretion
[[cite:PUB00014294]].

1620

Functional Annotations (from Homology via BLASTp):

- GO terms associated with the homolog:
- GO term1: 0005515 • definition: Binding to a protein.

Answer:

As a senior systems biologist, my analysis of the provided data leads to the following conclusion about the protein's function. Summary of Function: Based on the provided information, this protein is the **Translocated Intimin Receptor (Tir)**, a key multi-domain effector protein from attaching and effacing (A/E) bacterial pathogens like enteropathogenic *E. coli* (EPEC). Its primary function is to act as a molecular bridge that subverts host cell processes to facilitate bacterial adhesion and pathogenesis. After being injected directly into the host cell by the bacterium, Tir inserts into the host cell membrane and functions as a high-affinity receptor for the bacterial outer membrane adhesin, **intimin**. This specific interaction is the central event that triggers a cascade of host signaling events, leading to dramatic cytoskeletal rearrangements and the formation of an 'actin pedestal' structure beneath the bacterium.

Detailed Functional Analysis by Domain: The protein's overall function is accomplished through the coordinated action of its distinct domains:

1. **Tir Intimin-Binding Domain (IBD) (PF03549):** This is the core interaction domain. Its specific function is to **bind directly to intimin**. This binding event is essential for the intimate attachment of the bacterium to the host cell and for establishing the correct topology of the Tir protein within the host membrane.

2. **Tir C-terminal Domain (PF07489):** This domain is also crucial for host interaction. It is sufficient for **recognizing and binding intimin**, reinforcing the function of the IBD. Additionally, it has a secondary role in **binding to host cell beta-1 integrins**, suggesting it may modulate or interfere with native host cell adhesion and signaling pathways, although the precise role of this interaction is noted as unclear.

3. **Tir N-terminal Domain (PF07490):** This domain's function is related to the protein's own lifecycle and deployment. It is involved in ensuring the **stability of the Tir protein** and is necessary for its **secretion** from the bacterium into the host cell via the Type III secretion system.

The associated Gene Ontology term, **GO:0005515 (protein binding)**, serves as a high-level functional annotation that is strongly supported by the specific binding activities described for both the IBD (PF03549) and the C-terminal domain (PF07489).

Role in the Biological System: From a systems perspective, Tir is a masterful example of pathogenic manipulation. The pathogen does not rely on a pre-existing host receptor; instead, it injects its own custom receptor into the host. The process is as follows:

1. The A/E pathogen uses its Type III Secretion System to inject Tir into a mammalian intestinal cell.
2. The Tir protein, guided by its N- and C-terminal domains, inserts into the host cell membrane, exposing its central intimin-binding domain to the exterior.
3. Intimin on the bacterial surface binds with high affinity to the translocated Tir.
4. This binding event triggers host signaling pathways, leading to the recruitment and nucleation of actin filaments directly beneath the bacterium.
5. This results in the formation of an 'actin-rich pedestal' and the destruction of the local intestinal microvilli (the A/E lesion), which is the hallmark of this pathogenic mechanism and is critical for causing disease.

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674
 1675 In summary, Tir is a multi-functional protein that acts as a
 1676 central hub in bacterial pathogenesis. It ensures its own
 1677 stability and secretion, inserts into the host membrane, and
 1678 orchestrates the specific binding and signaling events required
 1679 to hijack the host's cytoskeleton for the bacterium's benefit.
 1680

1681 **Ground Truth:**
 1682 Multifunctional protein that is required for efficient pedestal
 1683 formation in host epithelial cells during infection. The
 1684 extracellular region acts as a receptor for bacterial intimin,
 1685 allowing the bacterium to attach tightly to the host-cell
 1686 surface. Simultaneously, the intracellular region initiates
 1687 a signaling cascade in the host cell, which leads to actin
 1688 polymerization and formation of actin pedestals at the sites of
 1689 bacterial adhesion (By similarity)

1690 **I.2.2 EXAMPLE 2: PATHWAY PREDICTION**

1691
 1692 The example of PDB ID: P78748
 1693

1694 You are a senior systems biologist. Analyze the input
 1695 information to answer the given question.
 1696

1697 **Question:**
 1698 What is the pathway of this protein?
 1699

1700 **Conserved Domains (from Pfam):**
 1701

- 1702 • **PF00215:** This family includes Orotidine 5'-phosphate
 1703 decarboxylase enzymes [ec:4.1.1.23] that are involved in the
 1704 final step of pyrimidine biosynthesis. The family also includes
 1705 enzymes such as hexulose-6-phosphate synthase. This family
 1706 appears to be distantly related to [pfam:PF00834].

1707 **Functional Annotations (from Homology via BLASTp):**
 1708

- 1709 - GO terms associated with the homolog:
 1710
- 1711 • **GO 0044205:** The chemical reactions and pathways resulting in
 1712 the formation of UMP, uridine monophosphate, starting with
 1713 the synthesis of (S)-dihydroorotate from bicarbonate; UMP
 1714 biosynthesis may either occur via reduction by quinone, NAD+
 1715 or oxygen.
- 1716 • **GO 0005829:** The part of the cytoplasm that does not contain
 1717 organelles but which does contain other particulate matter, such
 1718 as protein complexes.
- 1719 • **GO 0004590:** Catalysis of the reaction: H+ + orotidine
 1720 5'-phosphate = CO2 + UMP.
- 1721 • **GO 0006207:** The chemical reactions and pathways resulting in
 1722 the formation of pyrimidine nucleobases, 1,3-diazine, organic
 1723 nitrogenous bases, beginning with the synthesis of a pyrimidine
 1724 ring from simpler precursors.

1725 **Answer:**
 1726

1727 Based on a systems-level analysis of the provided data, the
 1728 protein is a key enzyme in the **de novo pyrimidine biosynthesis**
 1729 **pathway**.
 1730

1731 Here is a breakdown of the evidence:
 1732

- 1733 1. **Broad Pathway Identification:** The motif (PF00215) and
 1734 GO term GO:0006207 both explicitly place the protein within

1728
 1729 “pyrimidine biosynthesis.” This is the overarching metabolic
 1730 network responsible for creating pyrimidine nucleobases.
 1731 2. **Specific Sub-Pathway and Function:** The protein’s role is
 1732 further refined to the **UMP biosynthesis** sub-pathway (GO:0044205).
 1733 The motif describes its function as the ‘‘final step’’ in this
 1734 process.
 1735 3. **Molecular Mechanism:** Its precise enzymatic function is
 1736 Orotidine 5'-phosphate decarboxylase activity (GO:0004590),
 1737 catalyzing the conversion of orotidine 5'-phosphate into UMP.
 1738 This confirms its role as the terminal enzyme in the synthesis of
 1739 UMP, a foundational pyrimidine nucleotide.
 1740 4. **Cellular Location:** The pathway is localized to the **cytosol**
 1741 (GO:0005829).
 1742 Conclusion: This protein is an **Orotidine 5'-phosphate**
 1743 **decarboxylase** that executes the final, decisive step of the **UMP**
 1744 **biosynthesis pathway**. This pathway is a critical component of
 1745 the larger **pyrimidine biosynthetic network**, and it occurs within
 1746 the cytosol of the cell.
 1747
 1748
 1749
 1750
 1751 **Ground Truth:**
 1752 Pyrimidine metabolism; UMP biosynthesis via de novo pathway; UMP
 1753 from orotate: step 2/2

I.2.3 EXAMPLE 3: SUBCELLULAR LOCATION PREDICTION

1754 The example of PDB ID: P63171
 1755
 1756 You are a senior systems biologist. Analyze the input
 1757 information to answer the given question.
 1758
 1759 **Question:**
 1760 What is the subcellular location of this protein?
 1761
 1762 **Conserved Domains (from Pfam):**
 1763 • **PF03645:** Tctex-1 is a dynein light chain. It has been shown
 1764 that Tctex-1 can bind to the cytoplasmic tail of rhodopsin.
 1765 C-terminal rhodopsin mutations responsible for retinitis
 1766 pigmentosa inhibit this interaction.
 1767
 1768 **Answer:**
 1769 Based on the analysis of the provided information, the
 1770 subcellular location of Tctex-1 is the **cytoplasm**.
 1771 Here is the reasoning:
 1772 1. *Component of the Dynein Motor Complex:* The protein is
 1773 identified as a ‘‘dynein light chain.’’ Dynein is a motor
 1774 protein complex that functions within the cytoplasm, moving
 1775 cellular cargo along microtubule tracks.
 1776 2. *Specific Binding Partner:* The text explicitly states that
 1777 Tctex-1 binds to the cytoplasmic tail of rhodopsin. For this
 1778 interaction to occur, Tctex-1 must be located in the cytoplasm
 1779 where it can access this specific domain of rhodopsin.
 1780
 1781 **Ground Truth:**
 1782 Golgi apparatus, Cytoplasm, cytoskeleton, spindle

1782 **J LIMITATIONS**
1783

1784 While our context-driven approach demonstrates impressive performance in various bioinformatics
 1785 tasks, there are still significant challenges, particularly in predicting the effects of protein mutations.
 1786 One of the primary difficulties is that InterProScan and BLAST—two key bioinformatics tools that
 1787 we use for generating functional context—are not sensitive to subtle mutations at a single or few
 1788 amino acid positions. This is because both tools rely on sequence-level comparisons, and mutations
 1789 that alter a few amino acids may not result in any meaningful changes in the overall sequence
 1790 context, making the corresponding Pfam domains or GO terms identical before and after mutation.

1791 Thus, when we apply our method to predicting the effects of mutations, we encounter a situation
 1792 where the context generated for the wild-type and mutated proteins is essentially the same. This
 1793 leads to the limitation that our approach, at present, cannot effectively predict changes in the pro-
 1794 tein’s function or characteristics due to small mutations.

1795 Below are two examples demonstrating this limitation: one shows the context for a wild-type protein,
 1796 and the other for a mutated version of the same protein. The only difference between the two
 1797 sequences is the mutation at two amino acid positions, which we have highlighted in **red**. For clarity
 1798 and ease of comparison, we have provided only the Pfam domain and GO annotations (numbers)
 1799 rather than the complete context, which would otherwise be too long to display for these examples.
 1800

1801 **J.1 WILD-TYPE PROTEIN EXAMPLE**
18021803 **WT - Wild-Type**1805 **Conserved Domains (from Pfam) :**

- 1806 • **PF00732**

- 1807 • **PF05199**

1808 **Functional Annotations (from Homology via BLASTp) :**

- 1809 • **GO 0005737**

- 1810 • **GO 0005576**

- 1811 • **GO 0046562**

- 1812 • **GO 0050660**

- 1813 • **GO 0044550**

1814 **Sequence:**

1815 GIEASLLTDPKEVAGRTVDYIIAGGGLTGLTTAARLTENPDITVLVIESGSYES
 1816 DRGPPIIEDLNAYGDIFGSSVDHAYETVELATNNQTLALIRSGNGLGGSTLVNGGT
 1817 WTRPHKAQVDSWETVFGNEGWNWDSVAAYSLQAERARAPNAKQIAAGHYFNASC
 1818 HGINGTVHAGPRDTGDDYSPIVKALMSAVEDRGVPTKKDLGCGDPHGVSMPNT
 1819 LHEDQVRSDAAREWLDPNYQRPNLQVLTGQYVGKVLLSQNATTPRAVGVFGTH
 1820 KGNTHNVYAKHEVLLAAGSAVSPITLEYSGIGMKSILEPLGIDTVVDLPVGLNL
 1821 QDQTTSTVRSRITSAGAGQGQAAWFATFNETFGDYTEKAHELLNTKLEQWAEEA
 1822 VARGGFHNTTALLIQYENYRDWIVKDNVAYSELFLDTAGVASFDVWDLLPFTRG
 1823 YVHILDKDPYLRFAYDPQYFLNELDLLGQAAATQLARNISNSGAMQTYFAGET
 1824 IPGDNLAYDADLRRAWTEYIYPYNFRPNYHGVGTCSSMMPKEMGGVVVDNAARVYGVQ
 1825 GLRVIDGSIPPTQMSSHVMTFYAMALKIADAVLADYASMQ
 1826

1827 **J.2 MUTATED PROTEIN EXAMPLE**
18281829 **MUT1 - Mutated**1830 **Conserved Domains (from Pfam) :**

1831 [same as WT]

1836
 1837 **Functional Annotations (from Homology via BLASTp) :**
 1838 [same as WT]
 1839 **Sequence :**
 1840 GIEASLLTDPKEVAGRTVDYIIAGGLTGLTTAARLTENPDITVLVIESGSYES
 1841 DRGPPIIEDLNAYGDIFGSSVDHAYETVCLATNNQTALIRSGNGLGGSTLVNGGT
 1842 WTRPHKAQVDSWETVFGNEGWNWDSSVAAYSLSQAERARAPNAKQIAAGHYFNASC
 1843 HGINGTVHAGPRDTGDDYSPIVKALMSAVEDRGVPTKKDLGCGDPHGVSMFPNT
 1844 LHEDQVRSDAAREWLLPNYQRPNLQVLTGQYVGKVLLSQNATTPRAVGVFGLN
 1845 KGNTHNVYAKHEVLLAAGSAVSPITLEYSGIGMKSILEPLGIDTVVDPVGLNL
 1846 QDQTTSVRSRITSAGAGQGQAAWFATFNETFGDYTEKAHELLNTKLEQWAEEA
 1847 VARGGFHNTTALLIQYENYRDWIVKDNVAYSELFLDTAGEASFVDWDLPPFTRG
 1848 YVHILDKDPYLRHFAYDPQYFLNEDLLGQAAATQLARNISNSGAMQTYFAGET
 1849 IPGDNLAYDADLRAWTEYIPYNFRPNYHVGTCSSMMPKEMGGVVVDNAARVYGVQ
 1850 GLRVIDGSIPPTQMSSHVMVFYAMALKIADAVLADYASMQ

1850 In both cases, the context generated for the wild-type and mutated proteins is identical, as Inter-
 1851 ProScan and BLAST fail to identify significant differences caused by mutations at amino acid po-
 1852 sitions. This limitation underscores the challenge of predicting the effects of mutations based on
 1853 sequence context alone, as even small changes in the sequence can have profound effects on protein
 1854 function that are not captured by these tools.

1855 J.3 SUMMARY OF THE LIMITATION

1856 The core limitation of our current method is its inability to distinguish the functional consequences
 1857 of amino acid mutations, due to the lack of sensitivity of InterProScan and BLAST to small sequence
 1858 changes. This highlights the complexity of mutation prediction tasks, where even subtle changes to
 1859 the sequence can lead to vastly different biological outcomes.

1860 K COMPARISON OF MODEL PERFORMANCE ON THE SAME PROTEIN AND 1861 QUESTION

1862 In this section, we compare the performance of different models on the task of predicting the function
 1863 of the same protein, specifically protein **A7YWC4**, using the question: “What is the function of this
 1864 protein” The ground truth for this protein is as follows:

1865 Ground Truth

1866 **Ground Truth:** Essential for mitochondrial network organization, mitochondrial metabolism and
 1867 cell growth at the organism and cellular level. May play an important role in mitochondrial protein
 1868 synthesis, mitochondrial DNA replication, and contribute to nucleoid stability. Also involved in
 1869 mitochondrial-mediated antiviral immunity and the mitochondrial DNA damage response, promot-
 1870 ing signaling between damaged genomes and the mitochondrial membrane, which leads to activa-
 1871 tion of the integrated stress response (ISR).

1872 Below, we compare the answers generated by our method (using context), Interns1, Evolla, and
 1873 NatureLM.

1874 Our Method (Context-Driven, Gemini2.5Pro)

1875 **Answer:**

1876 As a senior systems biologist, my analysis of the provided information leads to the following con-
 1877 clusion about the function of this protein.

1878 **Executive Summary**

1879 This protein is **ATAD3 (ATPase family AAA domain-containing protein 3)**. It is a multi-
 1880 functional mitochondrial protein that acts as a central hub for organizing mitochondrial structure
 1881 and function. Its core activity is derived from its AAA ATPase domain, which uses the energy from
 1882 ATP hydrolysis to perform chaperone-like functions, likely remodeling protein complexes. This

1890

molecular activity is essential for maintaining the mitochondrial network's architecture, regulating mitochondrial metabolism, and ultimately, supporting cellular growth.

1891

Detailed Functional Analysis

1892

The protein's function can be broken down by integrating the motif data with the relevant Gene Ontology (GO) terms.

1893

1. Protein Identity and Core Mechanism

1894

- **Identity:** The presence of the **PF12037** domain specifically identifies this protein as **ATAD3**.

1895

- **Mechanism:** The protein contains a **PF00004** (AAA ATPase) domain. This domain confers two key molecular functions:

1896

- **ATP binding (GO:0005524):** As an ATPase, it must first bind ATP.

1897

- **ATPase activity (GO:0016887):** It hydrolyzes ATP to ADP to release energy. This energy powers its primary role as a **chaperone-like protein** that assists in the “assembly, operation, or disassembly of protein complexes.”

1898

2. Biological Processes

1899

The energy-dependent chaperone activity of ATAD3 is leveraged to control several critical cellular processes:

1900

- **Mitochondrion organization (GO:0007005):** This is a primary function. The PF12037 motif description explicitly states that ATAD3 is “essential for mitochondrial network organisation.” This involves managing the shape, size, and distribution of mitochondria within the cell.

1901

- **Regulation of cell growth (GO:0001558):** The motif information directly links ATAD3’s function to being “essential for... cell growth at organism and cellular level.” Healthy mitochondrial function is a prerequisite for cell proliferation.

1902

- **Mitochondrial Metabolism and Protein Synthesis:** The PF12037 description notes its importance in “mitochondrial metabolism” and a potential role in “mitochondrial protein synthesis.” This is consistent with its organizational role, as proper structure is required for efficient metabolic function and positioning of ribosomes. Its association with the **mitochondrial nucleoid (GO:0042645)**, where mitochondrial DNA resides, would be critical for these processes.

1903

3. Cellular Component (Localization)

1904

The protein’s functions are tied to its specific location within the cell:

1905

- **Mitochondrion (GO:0005739):** The protein is fundamentally mitochondrial.

1906

- **Mitochondrial Membranes:** The PF12037 domain “interacts with the inner surface of the outer mitochondrial membrane.” This positions the protein at a critical interface. Given its role in organizing the entire organelle, it is highly likely to also associate with the **mitochondrial inner membrane (GO:0005743)**.

1907

- **Membrane Contact Sites (GO:0044233):** ATAD3’s role in organizing the mitochondrial network and its localization at the membrane interface strongly suggest its presence at **endoplasmic reticulum-mitochondrion membrane contact sites**. These sites are crucial for inter-organelle communication and lipid exchange, which are vital for mitochondrial biogenesis and dynamics.

1908

Integrated Synopsis

1909

ATAD3 functions as a molecular machine anchored to the mitochondrial membranes. It uses the energy from ATP hydrolysis to remodel protein structures, thereby orchestrating the complex architecture of the mitochondrial network. By controlling the physical organization of the mitochondria, including its association with the nucleoid and contact sites with the ER, ATAD3 governs essential downstream processes. This structural regulation is indispensable for efficient mitochondrial metabolism, protein synthesis, and, consequently, for the overall growth and viability of the cell. Any disruption in ATAD3 function would lead to a collapse of the mitochondrial network, metabolic failure, and impaired cell growth.

1910

Score: 100

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

Interns1 (Sequence Input)

Answer:

Essential for mitochondrial network organization, mitochondrial metabolism and cell growth at organism and cellular level. May play an important role in mitochondrial protein synthesis. May

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

1998

1999

also participate in mitochondrial DNA replication. May bind to mitochondrial DNA D-loops and contribute to nucleoid stability. Required for enhanced channeling of cholesterol for hormone-dependent steroidogenesis. Involved in mitochondrial-mediated antiviral innate immunity. Also involved in the mitochondrial DNA damage response by promoting signaling between damaged genomes and the mitochondrial membrane, leading to activation of the integrated stress response (ISR).

LLM Answer: The protein described by the provided amino acid sequence is likely an **ATPase** involved in energy-dependent cellular processes. Here's a detailed breakdown of its inferred function:

1. ATPase Domain:

- The sequence contains a **Walker A motif** (GXXXXGKT), a hallmark of ATP-binding sites, and a potential **Walker B motif** (DExxG), which is critical for ATP hydrolysis. These motifs indicate the protein uses ATP hydrolysis to drive conformational changes or mechanical work.

2. Structural Features:

- **Coiled-coil regions:** Suggest oligomerization or protein-protein interactions, common in ATPases that form multimeric complexes (e.g., hexamers).
- **N-terminal proline/glycine-rich region:** May serve as a flexible linker or binding site for regulatory proteins or substrates.

3. Functional Inference:

- The protein is likely involved in **energy-dependent processes** such as:
 - **Protein remodeling/disaggregation** (e.g., chaperones like Hsp104).
 - **Vesicle trafficking** (e.g., NSF, which disassembles SNARE complexes).
 - **DNA repair/replication** (e.g., helicases that unwind DNA).
 - **Proteolysis** (e.g., subunits of ATP-dependent proteases like Lon or Clp).

4. Key Motifs and Comparisons:

- The Walker A/B motifs align with **AAA+ ATPases**, a family involved in diverse functions like protein degradation, membrane fusion, and DNA metabolism.
- The absence of protease-specific motifs (e.g., Ser-Lys dyad in Lon) or helicase signatures (e.g., DExD in DEAD-box proteins) makes these less likely, though not excluded.

Conclusion: The protein is most likely a **AAA+ ATPase** functioning in **protein complex remodeling or vesicle trafficking**, with roles in processes such as SNARE complex disassembly (like NSF) or stress response (like Hsp104). Further experimental validation (e.g., domain mapping, interaction studies) would refine its specific role.

Score: 30

Evolla (Sequence Input)

Answer:

This protein, which is associated with the mitochondrial membrane in the organism *Oryza sativa* (rice), serves a multifaceted role primarily within the mitochondria. It is known as the Mitochondrial import receptor for the endoplasmic reticulum (MIA40), and its functions are intricately tied to the protein import and quality control processes within the mitochondria.

The primary function of this protein is to act as a receptor for the mitochondrial presequence translocase (MPT), which is crucial for the import of proteins into the mitochondrial matrix. This process is essential for the proper functioning of the mitochondria, as it ensures that the necessary proteins are correctly targeted and translocated into the organelle.

Additionally, the protein is involved in the translocation of proteins into the mitochondrial matrix. It is likely to recognize mitochondrial presequences that are less than 10 amino acids long, which are critical for the proper targeting and import of these proteins. This suggests a specialized role in the import of a subset of mitochondrial proteins.

The protein also plays a role in regulating the mitochondrial permeability transition pore (mPTP), which is a non-selective pore that can lead to mitochondrial swelling and cell death if activated. The regulation of this pore is vital for maintaining mitochondrial homeostasis and preventing apoptosis.

1998
 1999
 2000
 2001
 2002
 2003
 2004
 2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024
 2025
 2026
 2027
 2028
 2029
 2030
 2031
 2032
 2033
 2034
 2035
 2036
 2037
 2038
 2039
 2040
 2041
 2042
 2043
 2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051

In terms of its interaction with other proteins, the protein forms homohexamers, and it is part of a complex that includes the MICOS complex subunit MIC10. This complex is involved in the regulation of the mitochondrial membrane potential and the opening of the mPTP. The protein's interaction with the mPTP is also noted to be indirect, suggesting a complex regulatory network within the mitochondria.

Moreover, the protein interacts with the mitochondrial presequence translocase, which is a heterohexamer composed of subunits TIM10, TIM22, and TIM23. This interaction is essential for the import of proteins into the mitochondrial matrix.

The protein's involvement in various cellular processes is further highlighted by its association with the biological process of protein import into the mitochondrial matrix, as well as its role in the regulation of mitochondrial membrane potential and the opening of the mPTP. These functions are integral to maintaining mitochondrial homeostasis and preventing cell death, making the protein a key player in cellular survival and metabolic regulation.

Score: 30

NatureLM (Sequence Input)

Answer:

Upon analysis of the given protein sequence, it appears that the primary function of this protein is RNA binding, and it could be implicated in the mRNA transport.

Score: 30

K.1 ANALYSIS OF MODEL DIFFERENCES

Our Method provided the most accurate and complete functional analysis, correctly identifying the protein as **ATAD3** and detailing its involvement in mitochondrial network organization, metabolism, and cell growth, which closely aligns with the ground truth. The score of 100 reflects the comprehensive nature of the answer and the high alignment with the expected protein function.

Interns1, relying solely on sequence input, inferred the protein to be an ATPase, which is a reasonable prediction given the presence of specific ATP-binding motifs. However, it failed to identify the specific protein (ATAD3) and did not connect its functions to the mitochondrial network organization, leading to a much lower score of 30.

Evolla also struggled with a correct protein identification, suggesting a mitochondrial import receptor for the endoplasmic reticulum (MIA40), which does not match the true function of ATAD3. This error resulted in a score of 20.

NatureLM provided a very generic answer, linking the protein to RNA binding and mRNA transport, which is not at all related to the actual function of ATAD3. This misinterpretation also earned a score of 20.

In conclusion, while all models gave some plausible biological functions, none of them fully captured the detailed and specific roles of ATAD3 within the mitochondrial network, as outlined in the ground truth. Our method, leveraging context, was able to provide the most accurate and thorough analysis of the protein's function, demonstrating the advantage of context-driven approaches over sequence-based models in protein functional prediction tasks.

L THE USE OF LARGE LANGUAGE MODELS (LLMs)

In this work, we employed a Large Language Model (LLM) as an assistive tool to enhance the quality of our writing. Specifically, the LLM was utilized to identify and correct grammatical errors, as well as to refine the phrasing and style of the text to improve clarity and readability. Additionally, the LLM assisted in formatting the prompts within the “QA examples” section to ensure consistency and alignment with our intended experimental setup.

In accordance with the ICLR policies on LLM usage, we disclose our use of the LLM in this section and affirm that the assistance provided by the LLM did not compromise the originality or integrity of our work. We take full responsibility for the content of this manuscript and have ensured that all information is accurate and appropriately referenced.

2052 M DETAILED COST ESTIMATION 2053

2054 To ensure transparency, we provide the basis for our cost estimations using publicly available on-
2055 demand pricing and API costs.
2056

2057 Settings 2058

- 2059 • **CPU Instance (Single):** An economical instance such as `c6a.xlarge` at **\$0.153 / hour**.
- 2060 • **CPU Instance (Batch):** A powerful instance like `c6a.24xlarge` at **\$2.7 (each is 1.35) / hour**.
- 2061 • **GPU Instance (Single A100):** A single A100 instance at **\$2.75 / hour**.
- 2062 • **LLM API Cost:** A single call to DeepSeek-V3 is estimated at **\$0.000446**.
2063

2064 Single-Sequence Inference 2065

- 2066 • **Our Method Cost: $\approx \$0.0030$**

- 2067 1. **CPU Cost:** $(\$0.153 / 3600 \text{ seconds}) * 60 \text{ seconds} \approx \0.00255 .
- 2068 2. **API Cost:** $\$0.000446$.
- 2069 3. **Total:** $\$0.00255 + \$0.000446 = \$0.002996 \approx \0.0030 .

- 2070 • **Evolla Cost: $\approx \$0.0690$**

- 2071 1. **GPU Cost:** $(\$2.75 / 3600 \text{ seconds}) * 90 \text{ seconds} \approx \$0.06875 \approx \$0.0690$.
2072

2073 **Batch Processing Inference** The per-sequence averages for batch processing were derived by
2074 modeling a large-scale run to capture amortization and throughput effects.
2075

- 2076 • **Our Method: $\approx \$0.00054$ per sequence**

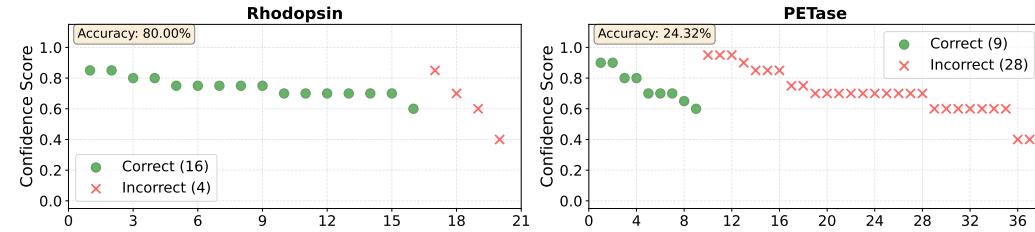
- 2077 1. **Throughput Model:** 2 powerful CPU machines process 1.12M sequences in 40 hours.
- 2078 2. **CPU Cost:** $(2 \text{ machines} * 40 \text{ hours} * \$1.35/\text{hour}) / 1.12\text{M seq} \approx \$0.000096/\text{seq}$.
- 2079 3. **API Cost:** $\$0.000446/\text{seq}$.
- 2080 4. **Total:** $\$0.000096 + \$0.000446 = \$0.00054/\text{seq}$.

- 2081 • **Evolla: $\approx \$0.0152$ per sequence**

- 2082 1. **Throughput Model:** 1 A100 machine (with 20x concurrency) processes ≈ 180 sequences per hour.
- 2083 2. **GPU Cost:** $(1 \text{ hour} * \$2.75/\text{hour}) / 180 \text{ seq} \approx \$0.0152/\text{seq}$.
2084

2085 N WET-LAB VALIDATION SAMPLES 2086

2087 N.1 PROMPT 2088



2097 Figure 13: Sample-level performance on classification accuracy of Intern-S1.
2098

2099 To evaluate the models’ performance on the novel protein classification task, we designed a stan-
2100 dardized prompt that frames the problem as a zero-shot binary classification. For each protein family
2101 under investigation (i.e., Rhodopsin or PETase), the model was tasked with determining whether the
2102 provided input protein belonged to the specified family or to a generic “Other” category. For the
2103 structure-aware model, Evolla, we first generated high-quality three-dimensional structure predic-
2104 tions for each novel sequence using the state-of-the-art AlphaFold3. It should be noted that the
2105 NatureLM model consistently failed to produce coherent or valid classifications across all samples
2106 in preliminary tests; it was therefore excluded from the final comparative analysis to maintain focus
2107 on viable models. Intern-S1’s results are shown in Figure 13.

2106

2107

2108

2109

2110

2111

2112

2113

2114

2115

Prompt for Novel Protein Sequence Classification

What is the category of this protein?

(1) Rhodopsin/PETase (2) Other

Please provide your answer in the following structured format: Category: [1, 2] Confidence score: [a score between 0.0 and 1.0]

N.2 SEQUENCE LIST

PET Samples

2119 >seq_83

MATLAPQADSQNPYERGPAPLSEITAETGPFDIGSVTVAGSGQGFNSGTVVYPTDSEGTGFAIAIMPGFLASQSDIAWYGPRLAS
QGFVVMTLNTNALWDPFDRSRQLAALTYLTAQSTVKNRIDPARLAVMGWSMGGGSLQSAASTPSLAAIPLAPWDLNVSSRITVP
TMIFGADGDTVASVDDALPFYNGLANAPDKSLIVLKTDHFATPKNTIAQYSVSWLKRFDVNDTRYDQFLCPTNDPHTVVFQITCP
L

2120 >seq_112

MWRRGTAAGAVGAVVLGALTVAAVQPASGADNPYQRGDPDTQSSVSASRGTFATAQVGVPAGNGFGGVIYHPTDTSQGTFGAVAV
PGYATWAAEAGWMGHWLSFGFVIGIDTNNRNDWDTARGTQLLAALDYLQTQSSVRDRVDPRTAV/MGHSMGGGAMYAAQLQR
PSLKTAVGLAEPSPSQNLTTTVPMLSGONDTTTPASIVNLVNGIPSSVEKAYLELTGAGHGFPTSNNSTMNRVIPWLKIFVDSDTR
YTQFLCPLSDNSNGIRTYQNTCPLVPSPVTTSTTSSTTSTTTTTSVPVGACSATYRTTNAWSGGYQGEVTVAAGDSAVNGWTRV
WTLDSGQTISQVWNGTLSATGSAVSRNASYNGSLRGPTSTTGFIAAGGAPSTVLSCTSP

2121 >seq_673

MPRTAVSAKLSGSGQPRPRTHGSLGRLFRRSTTTLVAVTAVVASVISTAPASAADNPYERGPNPQTQASIAASRGTFATAEVSGGGNGFA
AAKIVYPTDTSQGTFGAVAVPGYATWAAEAGWMGPWLASFQFVIGIDTNSRNDWDTARGTQLLAALDYLQTQSSVRDRVDA
MGHSMGGGAAISALRRPSLKAAVGLAPFSPSQSLNNMRVPTMLLAGQNDGTTT PASILNHYNAIPAASEKAYLELAGADHGFP
VMMRKVIPWLKIFVQDFTRYQQFLCPLVWDWTGITSYQSSCPLPGGSTGTTYNLVNSSGKCADVNAAQSQVNGAALLWTCHTNPQRW
TETSAKERVYDNPKCMDAGAAQGQTRVTINSCSGGGSQKWTNVANGTVNDLSGRCLGAAGGATTNDNPLVVASCDGSPGQQWTKQ
S

2122 >seq_799

MPYKETDPQTSGNFRAGRLTPRTGLLGRIAVATAALFATAAVVAVRPALAADNPYQRGPNTAASVAASRGTFATTQVNVPAGNG
FFGGVIVYPTDTSQGTFGAVAVPGYATWAAEAGWMGPWLASFQFVIGIDTNSRNDWDTARGTQLLAALDYLQTQSSVRDRVDA
AAVMGHSMGGGAMYAAQLRPSLKAAVGLAPFSPSQSLNNMRVPTMLLAGQNDGTTT PASILNHYNAIPAASEKAYLELAGADHGFP
TSNNPTMMRNPWLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNSSGKCADVNAAQSQVNGAALLWTCHTNPQRW
CNGTGAVGAYAQFTVNTSAAGTAALGVRFANGASGARSANLVNGTTATVFSFESTGAWSTVTKNLTASLNAGSNTRLSPTVADGL
PNIDYLTVGAA

2123 >seq_1932

MFKLATAGIAAAVGMTMTVTAGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFGNGGRIYYPDTSQGTFGAIASPGYTA
LFASELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDSRSLAVGHSMSMGGGALSAAMRRASLKA
VVG ITPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNSSGKCADVNAAQSQVNGAALLWTCHTNPQRW
LLDNGSVVYTRSTCPPLPTDPTPTPPTGTPPTPAPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTP
NRQLQVYGNMCLDAAGSGNGAAVQIYSCHNQTNQWNINSNTISGVQSGRCLDVWSTNSGAQIQLYDCHGQTNQQFRLTPLA

2124 >seq_2113

MRGHGYHPPSGNTRADRVSFRGRFLKLAAGIAAAVGLLITVATGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG
GGRIIYPTDTSQGTFGAIASPGYTAFLSAELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDPTRLA
GHSMGGGAAISALRRPSLKAAVGIAPIYSPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNS
VMVRTMLPFVKMFIDNDARYSQQFLCPLLDSSGVVYTRSTCPPLPSTPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTP
TRVQLYGCNQTNQAWTYTSNKQLQVYGMCLDAAGTANGAAVQIYCHSQTNQWNVNSNGTISNVQSGRCLDAWSTANGA
QDCHGQTNQQFRLAALAQ

2125 >seq_2238

MRGHANHPPSENTRADRMSPRFRGRFLKLAAGIAAAVGMTMTVATGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG
FNGGRIYPTDTSQGTFGAIASPGYTAFLSAELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDPTRLA
VAGHSMGGGAAISALRRPSLKAAVGIAPIYSPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNS
SNPVLVRTMLPFVKMFIDNDARYSQQFLCPLLDSSGVVYTRSTCPPLPSTPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTP
RTNGTLYDCNRQINQWTYANKLQVYGMCLDAAGTANGAAVQIYCHSQTNQWNVNSNGTISNVQSGRCLDAWSTANGA
QIQLYDCHGQTNQQFRLVSLA

2126 >seq_2251

MSPRFRGRFLKLAAGIAAAVGMTMTVATGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFGNGGRIYYPDTSQGTFGAI
AISPGYTAFLSAELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDPTRLA
SLKAAGVIAPIYSPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNSNGTISNVQSGRCLDAWSTANGA
QIQLYDCHGQTNQQFRLVSLA

2127 >seq_2274

MRGHGHPPSGNSHTDRVSFRGRFLKLAAGIAAAVGLLITVATGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG
GGRIIYPTDTSQGTFGAIASPGYTAFLSAELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDPTRLA
GHSMGGGAAISALRRPSLKAAVGIAPIYSPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNS
VMVRTMLPFVKMFIDNDARYSQQFLCPLLDSSGVVYTRSTCPPLPSTPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTP
LYGCNQTNQAWTYTSNKQLQVYGMCLDAAGTANGAAVQIYCHSQTNQWNVNSNGTISNVQSGRCLDAWSTANGA
QIQLYDCHGQTNQQFRLAALAR

2128 >seq_2320

MRNDAHHPPSGPTATEPVRPDRRGRFLKLAAGVLAAGVGLITMTTGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG
NGGRIYPTDTSQGTFGAIASPGYTAFLSAELAWMGPWLASHGFVIGIETNSRNDFTARGTQLLAALDYLQTQSSVRDRVDPTRLA
AGHSMGGGAAISALRRPSLKAAVGIAPIYSPSSNLLANDRVPMLKIFDSDTRYTQFLCPLSNWSGIRAYQSSCPLPGGSTGTTYNLVNS
VMVRTMLPFVKMFIDNDARYSQQFLCPLLDSSGVVYTRSTCPPLPSTPTPPTPPTPPTPPTPPTPPTPPTPPTPPTPPTP
GTRVQLYGCNQTNQAWTYTSNKQLQVYGMCLDAAGTANGAAVQIYCHSQTNQWNVNSNGTISNVQSGRCLDAWSTANGA
QIQLYDCHGQTNQQFRLTPIA

2129 >seq_2498

MRGHTPHPTPSGDTDHARRSPFRGRFLTAAAGIAAVGLFAMVTTGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG
MRGHTPHPTPSGDTDHARRSPFRGRFLTAAAGIAAVGLFAMVTTGSASAADNPYQRGPDPRTSNTAVNGPFANTSVSVPSTGFG

2160
 2161
 2162
 2163
 2164
 2165
 2166
 2167
 2168
 2169
 2170
 2171
 2172
 2173
 2174
 2175
 2176
 2177
 2178
 2179
 2180
 2181
 2182
 2183
 2184
 2185
 2186
 2187
 2188
 2189
 2190
 2191
 2192
 2193
 2194
 2195
 2196
 2197
 2198
 2199
 2200
 2201
 2202
 2203
 2204
 2205
 2206
 2207
 2208
 2209
 2210
 2211
 2212
 2213

GGRIIYPTDTSQTFGAIAISPGYTAFLSAELAWMGPWLASHGFVVIGETNSRNDFTARGTQLLAALDYLTQQSPVRDRVDPNRLAVS
 GHSMGGGALSAAIRRSSLAKAVVGIAPYSPSSNLANDRVPTMIFSGQADTVTPTPSYATGLYNSLPATTESVYLEVAGADHGFMVGRSNP
 MIRTMLPFVKMFIDNDARYSQFLCPPLDSSGVTVYRSTCPPLPPATGTPTPSGNPTTPTPSTPTAPPGSASEIVGAQSNRCIDAPSRN
 NGTRVQLYDCNRQSNQLWTVTSKQKQLQVSGNMCDAAGSGNGAAVQIYSCHSQTQNQWNVNSNGTISGVQSGRCLDVWSTANGAQV
 QLYDCHGQTNNQQFRLVSR
 >seq_3000
 MRGHTPHPPPSRSDTHTDRVSLGLRGLLTLAAAGVAVTGLFTMAVTGSASAADNPYQRGPDPTRTSVAVNGPFANTSVSVPYG
 NGGRIIYPTDTSQTFGAIAISPGYTAFLSAELAWMGPWLASHGFVVIGETNSRNDFTARGTQLLAALDYLTQQSPVRDRVDPNRLAV
 GHSMGGGALSAAIRRSSLAKAVVGIAPYSPSSNLANDRVPTMIFSGQADTVTPTPSYATGLYNSLPATTESVYLEVAGADHGFMVGRSNP
 VMVRTMLPFVKMFIDNDTRYSQFLCPPLDSSGVTVRSTCPPLSDPTPTPTPPTVPPGSASEIVGTQSGRCIDVPNASRTNGR
 TRVQLYDCNKNSQNAWTVTSKQRLVYGDMLDAAGSGNGAAVQIYSCHSQTQNQWNVNSNGTISGVQSGRCLDVWSTNGAQV
 YDCHGQTNNQQFRLVSR
 >seq_3002
 MREPADHLPAGATPADRASPGLRGLFKLAAAGIAAAVGLLISITLATGSASAADNPYQRGPDPTRTSVAVNGPFANTSVSVPYG
 GGRIIYPTDTSQTFGAIAISPGYTAFLSAELAWMGPWLASHGFVVIGETNSRNDFTARGTQLLAALDYLTQQSPVRDRVDPNRLAV
 GHSMGGGALSAAIRRSSLAKAVVGIAPYSPSSNLANDRVPTMIFSGQADTVTPTPSYATGLYNSLPATTESVYLEVAGADHGFMVGRSNP
 VMVRTMLPFVKMFIDNDTRYSQFLCPPLDSSGVTVRSTCPPLSDPTPTPTPPTVPPGSASEIVGTQSGRCIDVPNASRTNGR
 CHGQTNNQQFRLVSLG
 >seq_3008
 MRGHADHPLPSGNADHRVSPRSRGRGLFKLAAAGIAAAVGLFTITVATGSASAADNPYQRGPDPTRTSVAE
 NGPFANTSVSVPYG
 NGGRIIYPTDTSQTFGAIAISPGYTAFLSAELAWMGPWLASHGFVVIGETNSRNDFTARGTQLLAALDYLTQQSPVRDRVDPNRLAV
 GHSMGGGALSAAIRRSSLAKAVVGIAPYSPSSNLANDRVPTMIFSGQADTVTPTPSYATGLYNSLPATTESVYLEVAGADHGFMVGRSNP
 VMVRTMLPFVKMFIDNDARYSQFLCPPLDSSGVTVRSTCPPLSDPTPTPTPPTVPPGSASEIVGTQSGRCIDVPNASRTNGR
 YDCHGQTNNQQFRLVSLAR
 >seq_3382
 MRGPADHPPAGNTRADRVPSPGRGLFLTLAAAGIAAVGLFTITVATGSASAADNPYQRGPDPTRTSVAE
 NGPFANTSVSVPYG
 GGRIIYPTDTSQTFGAIAISPGYTAFLSAELAWMGPWLASHGFVVIGETNSRNDFTARGTQLLAALDYLTQQSPVRDRVDPNRLAV
 GHSMGGGALSAAIRRSSLAKAVVGIAPYSPSSNLANDRVPTMIFSGQADTVTPTPSYATGLYNSLPATTESVYLEVAGADHGFMVGRSNP
 VMVRTMLPFVKMFIDNDARYSQFLCPPLDSSGVTVRSTCPPLSDPTPTPTPPTVPPGSASEIVGTQSGRCIDVPNASRTNGR
 TRVQLYDCNKNSQNAWTVTSKQRLVYGDMLDAAGSGNGAAVQIYSCHSQTQNQWNVNSNGTISGVQSGRCLDAWSTNGAQV
 YDCHGQTNNQQFRLAALA
 >seq_3397
 MLRRAAIACAATLGLVLASTVAGPASAADNPQHGRGPNTQASVAASRGTFATSETSVGGNGFGGAKIYPTDTSQGKFAALAVSPGYT
 ATWAAEAGWMGHWLASFVFIIGDITISRTDWDVARGDQLLAALDYLTQRSVDRDVSARVGMGHSMGGGAIASGLKRPSLKV
 GNAPYASQNLNTNMVRPVALLAGQRDTVTPTATTNSYNNQVATTEKMYELTNAHSPFSTSNSVMTRKVIPWMKIFLDQDTRYQOFL
 CPLLDNTGIRTYRSCPLIGDPMPADTYRLAGAASGKCADVEGRSQANGAKLLWTCNGQTNRWTQATGELRVYDNPCKVEAGTS
 AQGTRVTINACNGGNSQKWTNVANGTVTNAVGRCLEDAAGTADNSPIVVNTCTGAGSNWNSKRS
 >seq_3423
 MPERVTLRPHPRRNAPLPGHRTNGPVRRLAVTIAVGLLVTAVAAPVASAADNPYQRGPAPTLASVAATRGTFATAQMTVSPGNF
 NGGTIYPTDTSGLTGVAVAVPGYTAFADEEAWMGHWLASFVVIIGVETNSRTDFDTARGTQLLAALDYLTQKSPVTRVDPNRLS
 VIGHSMGGGALVAERPRSLKATVGLAPFKPSGNLAADRVPIMMGGTADTTVTPSYLDGLHATLPASTENAYLQLTGANHLFPTKPN
 TLEMRLLIPWLKTFVDSDTRYQFQFLCPGPADPTGIAISYRGTCPYLPA
 >seq_3565
 MPEEHHTQPARVSGDWGIGASRRGLSAGPVRRIAVVAVLGLMIAAVVAGVAPASAADNPYQRGPDPTRLAGVAATRGTFATAQMSVPPG
 NGFNGGMIYPTDTSGLTGVAVAVPGYTAFADEEAWMGHWLASFVVIIGVETNSRTDFDTARGVQLLAGLDYLTQKSPVNRVDP
 NRSVLSHGMMGGGALVAERPRSLKAAVGLAPFKPSGNLAADRVPIMMGGTADTTVTPSYLDGLHATLPASTENAYLQLTGANHLFPTKPN
 PTKPNTLETRLIPWLKTFVNDNTTRYTQFLCPTRADPTGTSKYRSTCPYAPAGRSVSPSTADLSSL
 >seq_3627
 MPVVAHAAHHASGESAAARPARGPWRRTAVLSAALLGLSALTTASAAGTPTPPTPATAARVQAPAPAGSPYQRGPDPTRASVSASR
 GTFATAQLSVPPNGFNGGTTIYPTDTSGLTGVAVAVPGYTAFADEEAWMGHWLASFVVIIGVETNSRTDFDTARGVQLLAGLDYLTQKSPVDRP
 LTQKSPVDRPVPALSFGVVIIGVETNSRTDFDTARGVQLLAGLDYLTQKSPVNRVDP
 AYLQLTGANHLFATRPNLETRLIPWLKTFVNDNTTRYTQFLCPTRADPTGTSKYRSTCPYAPAGRSVSPSTADLSSL
 >seq_3633
 MANPYERGPNTDALLEASSGFSVSEENNSRLSASGFGGGTTIYPTDTSGLTGVAVAVPGYTAFADEEAWMGHWLASFVVIIGVETNSRTDFDTARGVQLLAGLDYLTQKSPVDRP
 PDSRAEQLNAALNMHRNASSRVSRLAVMGHSMGGGTPRLSAAKIPLTWHLNKNRSSVTPVTLIIGADLDTIAPVAT
 HAKPFYNSLSSKISAYLELDGATHFAPNIPKIIKGYSVAVLKRFDVNDNTTRYTQFLCPGRDGLFGEVEEYCSTCP
 >seq_3642
 MSAPTSPPITSTGLGEKISRRRGWRAKAAGVVAALALTTAVAAPAPAAANPYERGPDPDTSSIEATRGSFATSTDTVSRALAVSGFGGGT
 IYPTTITAGTGAISIAPGFTATQSSIAWLGPRLSAQGFVVFIDTLTTSDQPDSDRGRQLLASLDLTQSSVRSRDSRLLGVGHSMGGG
 TLEAARSRSPSLQAAVPLTGWNLTKNWSTLRVPTLVVGAQSDTIAVASHISPFYTSPLSTLDRAYELRGASHFAPNNSNTIAKYTLSW
 KRFIDNDTRYEQQFLCPITSLSIISDYRGNCPHNG
 >seq_3679
 MGGGVAAVATPALADEIIGQAPTSNITGNGFSFATTSESSISLVSFGFGGRVYYPYATGRYPVIAISPGFTATWSSLAWIGPRLSSWGFVV
 GIETNSIYDQPSRGNQNLAAALNWAVNSSTA VRSDVGSRRGVAGHSMGGGTLLEALAAADTSGLVAGVPLAPWNTDKIWSNVSEPV
 LIVGGQADTIAVATHSPVFPNTLAGPKTYVETGASHFPQTNTATTSRALVSWFKRWLQNQDSRFTPYTCFGGLAVSDFRSNAC
 >seq_3712
 MATAFVATALLAGGTGTASPAFADEIIGQAPTSNITGNGFSVSTYNISSLVSFGFGGVYYPYATGRYPVVAVSPGFTARWSSISW
 LASWGFVVVGIETNGLYDQPSRGSQLLAALNWAVNSAPSATSRADGSRRGVAGHSMGGGTLLEALAAADTTGVNKAGVPIAPWNSD
 TWSGINEAVQVGGESDTIASVGSHSIFYNSVSPKSVELNGASHFPQSTNSTLSRALVSWFKRWLQNQDARFTPTCGYGGTAVSDFR
 NAC
 >seq_3729
 MIATTFVATALAGGTGTASPAFADEIIGQAPTSNITGNGFSVSTYNISSLVSFGFGGVYYPYATGRYPVVAVSPGFTARWSSISW
 PRLASWGFVVVGIETNSVYDQPSRGSQLLAALNWAVNSAPSATSRADGSRRGVAGHSMGGGTLLEALAAADTTGVNKAGVPIAPWNS
 DKTWNVNEPVQIIGGESDTAVGVSNSVSPKSVELNGASHFPQSTNSTLSRALVSWFKRWLQNQDARFTPTCGYGGTAVSDFR
 DFRTNAC
 >seq_3982
 MQPQQKARRWFTRSSAVALVVLALAAGGLAATAAPAYADQVGQAPTAANITDGFSFATASAPITNTQGFGGGTVYYPYATGRYPVVA
 VVPGFVSTWSQISWLGPGRVAVSGWGVVVGADTTSGFDSPSQRADDELLAALNWAVNSAPAAVRGKVDGTRRGVAGWSMGGGTLLEAL
 KDTTGTVKAGVPLAPWDIGQDFSKVTPKPVFIVGAQNDTIAPPAQHAFVYNAAGPKSYLELAGDHFFPTTANPTVSRAMVSWLKR
 SSDDRFTPTCGFAGAAVSAFRSTAC
 >seq_4116
 MRAAATGLALALAVTAGATAVATPALADQOQAPTSNITDGFSFATVSSA VSGQTGFGGGTVYYPYATGRYPVVA
 WMGPRLASWGYVVGVNPNSVFDLPSRSQSQLLAALNWAVGSAPS VQSRADGTRGVGVYGS
 SPWDTNQNWSTVRQPEFIVGAQNDTAVPPAQAQHAFVYNAAGPKSYLELAGDHFFPTTANPTVSRAMVSWLKR
 TGSAISASMSNAC >seq_4359
 MTTPISQLRWWALRLTRAATALAAALALVGTAPAFADVEVGQAPTAANIVGDGSFATTSSA
 TGTQGFGGGTVYYPYATGRYPVVA
 APGFLSDWNSLNWLGSRSVAVSGFVVGVNTNSGFDPPSRQDQLLA
 ALNWAVGSAPAVRDKVGDTRGVAGWSMGGGTLLEALAK

2214 DTTGTVKAGVPLAPWHSDDKTWEQVSEPVFLVGGQNDTVAAPSSHAIPFYNSLGGQKSYLEAGQAHFFPTTSNPTVSRAIVSWFKRFVSS
 2215 DNRFPTFCYGTGSAVSNFRSSAC
 2216 >seq_5876
 2217 MAVGSMILLSMAAQAVVVFEEFTSTGLGKFTAAGSVVTSSGAARLDGCGYCTDGSITSTAISTVDTGLRLSFDRVTSGLDSEAGIAEF
 2218 STNGSTYTAVESIRTAISGRVTNLPTSAENQSLRLRFRINASLSETTYVDNIRLEGTSGGGTTNPFEKGPDTKTMLEASTGPFTYTT
 2219 IYRKVDPNRLGVGMWSMGGGTLIAARDNPTLKAAPFAPWNSSTNFSTSVPTLIAICESDSTAPVNSHASPFYNSLPSTTKAYLEMNN
 2220 GSHSCANSGNNSNAGLIGKYGVMKRFMDNDTRFSYLCGAPHQADLSLTADIERYRENCPY >seq_5913
 2221 MNFPRASRMLMQAAVGLGMAVSAATAATQNPYARGPNPTAASLEASAGPFTVRSFTVSRPSCYGAFTVYYPNAGGTVGAIAPVGYTA
 2222 RQSSIKWWGPRLASHGFVITIDTNSTLDQSSRSSQDMAALRQVSLNGTSSPIYGKVDTARMGVGMWSMGGGSLISAANPNSLKA
 2223 AAPQAPWDSSTNFSSVTPTLIFACENDSIAPVNSALPIYDMSMSRNAKQFLEINGGSHSCANSGNNSNQALIGKKGVAWMKRFMDNDTRY
 2224 STFACEPNTRVSDFRTANCS
 2225 >seq_6199
 2226 MPPDCVLPRLLAAAALLASATLVLPSAAAQTNPYQRGPDTTRDLEDSRGPFYASTNVRSPSGYAGTIYYPDVSISVGAVAVVPGYL
 2227 ARQSSIRWWGPRLASHGFVITIDTNSTLDQSSRSSQDMAALRQVSLNGTSSPIYGKVDTARMGVGMWSMGGGTLISARDNPSLK
 2228 AAVPFAPWHTNANFSGVQVPTLVIACENDTVAPIRHASSFYNSFSSSLAKAYLEINNGSHTCANTGNSNQALIGKKGVAWMKRFMDNDTRY
 2229 RYSPFLCGAPHQADLRSRSLSEYRENCPY
 2230 >seq_6214
 2231 MFGKLPFARASLAVGALLSAAAQAQTPYQRGPDTVSSLEATRGPSTSSFTVSRPSCYGAFTVYYPNAGGTVGAIAPVGYTARQS
 2232 SINWWGPRLASHGFVITIDTNSTLDQSSRSSQDMAALRQVSLNGTSSPIYGKVDTARMGVGMWSMGGGSLISAANPNSLRAAQP
 2233 APWAQESFSSTVPTLIVCENDSIAPVNSHSSFPFYNMQTRNKKANLNGHSCANSGNNSDAGLIGKKGVAWMKRFMDNDTRYSKFL
 2234 CGAEHQADLSKRAVEAYKENCOPY
 2235 >seq_6358
 2236 MAGAAGWPPTASADNPYQRGPAPTAASIAADRGPFATAQVTPRNGFGGVYIYPDTSQGTGAIIAIPGFTAWSYYAWQGPR
 2237 LASPGFVVVIGETNTLNDYADARAQLLAALDYLVNSSVDRDRNRLAVGGHSMGGGILAANQRPSSLLTAVQAPVYVPGDLS
 2238 GIKSPTIIFAGQVGDGTVPQYAQANYNTIPGSVERAYVEIANEGHFPAGGGGNGSAFARTMMVWLKFIDRDRTRYAPFLCPSTLNAN
 2239 GISKYMACSPCLDPPGGGPTASPTSTPTSTPTGTPPAGGTSALRGVASGRCLDNGAQSQANGATVLIWDCNGQNNQKWTSTS
 2240 AGELRVYGNKCLDNGGTLADSGAVIWIWDCNGQSNQKWLRSNDSGTSITAVQANKLDSVSGNGTANGTRVQIWSCTGGTNQKWTTRV
 2241 >seq_6382
 2242 MHPNRTVRLRAALIAVIGLLAALVSTAPATAQANPYQRGPDTVSSVAAQRGFTATAEVTAPVNGFNGKIVYYPDTSQGTWAAIAA
 2243 VPGYTAWRDNEGAWMGPWLASFVVGVIDGNTNSTLDWDTARGTQLLAALDYLQRSPVDRVDPNRLGVGHSMGGGAIASAERRP
 2244 ALKAAPVAPLAPPSQNLSTLRVPTMIVSARDSDVVTPLSVNDLYATMPAATPGAALLETTGGHGPWTGNSAATRRTIPWLKIFLDNDTR
 2245 YTQFLCPSLADLSGISRSRTKCPYGTALAAPETLQAVR
 2246 >seq_6547
 2247 MAGPTPTPRECTDVPHTPPTLRRMSSAATLLLVLLALAGGLAAPAAAHAAVHGPDTALLEASRGPYATAQIDVSSLSVSGFGGGTI
 2248 YYPPTTSEGTFFGVAAPGTYADAKSSLLAALDYLVNSSVDRDRNRLAVGGHSMGGGILAANQRPSSLLTAVQAPVYVPGDLS
 2249 GGGGTLLEAVADPRSLRAAPVPLTPWNLKDTWSGVQPTLIIAGDNDTVAAPVASHAEPFTSLPSSLDKAYLELNNAHFAPNTTNTTIGKY
 2250 AVAWMKRFVDDDTTRYGRFLCPGSSGLVEEYRSTCPF
 2251 >seq_6679
 2252 MSSAVTLLLVLLALAGGLAAPAAAHAAVHGPDTALLEASRGPYATAQIDVSSLSVSGFGGGTIYYPPTTSEGTFFGVAAPGTYADK
 2253 SSLAWLAARLASHGFVVFNIIDLRLDQPSRGRQLLAALDYLQRSSVRGRVDAEADPRSLRAAVPLTP
 2254 WNLDTWSGVQPTLIIAGDNDTVAAPVASHAEPFTSLPSSLDKAYLELNNAHFAPNTTNTTIGKYAVAWMKRFVDDDTTRYGRFLCP
 2255 PSSGLLVEEYRSTCPF
 2256 >seq_6781
 2257 MRRSPRNLLARLVTALALVAGAAATATASAQAADNPYERGPDTTESSIEARGPFAVSQTTVSSLLVSGFGGGTIYYPPTDTSQGTGAVAISP
 2258 GYTADQTSIAWGPRLASQGFVVFIDTITRDQPSRGRQLLAALDYLQRSSVRGRVDAEADPRSLRAAVPLTP
 2259 LAPWNLKNSFSSVQVPTIIGGESDTVAPVATHSEPFYTSLPSSLDKAYMELNNASHFPNTSNTMAKYMISWFKRFIDNDTRYDQFLCP
 2260 PNSGLAIEEYRSTCPY
 2261 >seq_6793
 2262 MKAAVVAATVGLTALVALVQJQPAQSSAAAANPYERGPAPTVSSVAAQRGFTATAEVTVPPNGFNGKIVYYPDTSQGTWGA
 2263 GYTARWSAEGAWMGPWLASFVVGVIDGNTSPNDYDTARGTQLLAALDYLTRKSPVHDRVDPNRLGVGHSMGGGAIASAERRPSL
 2264 AAVPLAPLAPPSQNLSSLRVPTMIMGARDGDTVPTSYLDGLYSGMMPAATQSAYIELTSGGGHGPWTGNSHVTTRAIPWLKIFLDNDTRYTQ
 2265 FLCPSPADSREVSRSQTKCPYVPPGTTTPSGEQIVGAASGRCLDVNNSSQTSQAAHPWDCADRGPKWARTDAGELRVYSNKLDA
 2266 EASGTSGTRVVIWACHGRQNQRWNVNANGTTISVPSGLCLDTYNGGTGNAQSVLGNCGNASTQRWTLR
 2267
 2268
 2269 Rhodopsin Samples
 2270
 2271 >seq_131
 2272 MGARSRTISILMAVQAVGALLVLLGERTLRAVTVTPQPTSTLHVDLGAAMVVLALSAAWALSAGAGARSSGAGARSSWWG
 2273 ALDPLLTPTILFVVAQQLNGIRDVGVALGVYALASAGVLFAVVQRDRDRTGGSRVPLGLGSAVIVPVGIVAFHQVAGIVGHPGP
 2274 VVITLTALVAAVAEFTWRRQLVAAVRLTAGFALVWLVAAL
 2275 >seq_735
 2276 MKNLSTNDYVGISFIIISIALAASTVFLLVERESMKPKWKLPSVVSAMVTGIAASHYYYMREIWWVSKKNPIHYRYIDWLTVPQIIEFYLI
 2277 LSVAKKIPFELFYKLASSIIMIVAGFLGESKVIDRNGQFIIIGTVAWLYIYLIEFYGEAKAFKEETDDESVKFAFDALKWITVGVWAIPIGY
 2278 LLEINKMNLNLGDLVNLKILFTLIIWYAGKQNGYIIPNIS
 2279 >seq_1245
 2280 MDLFIGSYVSVFQFISLFIQMYGYFFINVTPLPLKIALSIEFFSIVELLVYLWIGFNLSNLSSVMGKRYLDWFITTNFLVSVSLLMIY
 2281 FNQREERKSKYDGDKTPVQEQQNTNTLIERNLPKYAPLVVNNIMLLVGLGERGVLSKWISTPIGFIAFFLGFYHYHFFAKHSHSGKRF
 2282 YFITSIWAMYIAHHTFRDQPKNIAVNLDDLSKNAFGVFMVYMLHPNAFLI
 2283 >seq_1505
 2284 MFSWGDYRRGRTTNEQNQLINKDALGNKDALGLYSKIFACGGGVPFWISINWAAALMHLLNAFLTMLWDASDSKDNVFKLSEQYT
 2285 AWQPRANSTAECCGIDAFPISDEWIANATVTSSELSLWWLIIIVHFFLSFAQALAMAEDWDFSCCGYKCVRTSYIEEVDENEGTNSLRMIEYSI
 2286 SATLMQVSIALILSVWSALTMLGIVFLTIVTMLGLVAEKLKYDGLAWTVHFIGWLAMGAVAAILGRKFALNISMAPSGGGPVVFYV
 2287 IVVIAILYSGFGLVQFDFLMFTGDKKTAKSRRVKEVAMYCMLSLISKTLGLWIFSNALVGMAARN
 2288 >seq_1848
 2289 MLTGGDYLTVYVNLVLSFGTAVMFCAFVYFLQTMRSVNVQYQAGIAVSAIVVIGIAGYHYYRISDFGEVGMNEGORYADWLTIVPLIIEL
 2290 LIVLGVAKQDRRLSMLKLVPATILMVGVLGPGVEADNGAKWTFWVLAIPFILKLYGELQKAGQRETGAVASQIKNATYVLLTTW
 2291 MVYPIAYLFPIDGESEIETLRQVGYTFADIKGLYGLMLWIAKARTAEVTG
 2292 MTSNAQLFKRAGNQAVDNPAAVADIGMTHGSNWLWAVFCVFAFLSLYVAFPLTAERKSARLSKFAIASPLWISLIMTFAYFTMASL
 2293 GWTGQIAEFPNHASVNSSATPGIRQIYAKIYIWFCLCPWLLYYSPELATATTDVAEEQSRSSFESLHNLVQIAGVEFFVIGLLVGALIRSSY
 2294 KWGYWTMMAFAAILVAIVLFKHQQLVTLRVRGWKLIVLFLSTQLCVFLYFVCWGLSEGGNVIQPDSEVFYGLDLCVFAFLVPAFLSAFVSS
 2295 KGSLGFIPSLPMHGKNRNHGTDLEKQSNHEPVRSGETEVPEGVFGTTAPGTTIPTSTTHAANTTTA
 2296 >seq_2110
 2297 SQMSTFVVDLYKRGGEAIVKINPPTGADFHITSRGSDWLWAACFVFLLAIVFVLLMFRPKVNNRFLVYLTIAVNPWFMAIYFSIASNLW
 2298

2268
 2269
 2270
 2271
 2272
 2273
 2274
 2275
 2276
 2277
 2278
 2279
 2280
 2281
 2282
 2283
 2284
 2285
 2286
 2287
 2288
 2289
 2290
 2291
 2292
 2293
 2294
 2295
 2296
 2297
 2298
 2299
 2300
 2301
 2302
 2303
 2304
 2305
 2306
 2307
 2308
 2309
 2310
 2311
 2312
 2313
 2314
 2315
 2316
 2317
 2318
 2319
 2320
 2321

IPVRKYNHVRTSTQQQHPGVRQIFYARYVGWFMALPWPLIQASILGKTPVWHVAFNCTMGCVFSVCFLIAACVHSTYKWHGYFTIGCGA
 GIVSIIHSLMTTYYTLIKKCGDKEIKRCFLIVCPIVLYLVAWPVCFGITDGGNVLQDSEAFYGHIDLLLGIFPALLYPM
 >seq-3858
 MPAPETKQHHTFADLELESVRPRLRTGAATGATVETERLLNGDKSTMSEKASMTIIFCCYNFFCCWPFTPGLSGWCNPCLGDGFC
 CGDFFAKCFPRIRLGRKTWLRLHLLCFAIHTGWLVATIEVAWGKGPDMQVEILRKPKQWKGPKYGYEVAPSGIPVYRVDLIVCSFA
 LSALMHGMWVIDGLVVLVAHCFKRNNWGRVKGWLLWDRLDQCLCWWRCAVALPLPRSSQNIADPSCSHRWEASASVMFLALGIAI
 GVREQNALATIFFLSFTMWLGFTEMLSRPRTDMKSWEGDPPKALVEGTPLAFFVTKMRSYAWRMAPHVLGFFPYITAWVIIINNFVE
 QLNDLPEETRKLMPMEWWVVAIAGSVTISLFTFVQIRYQWIAPEHYWRTELWYCALSKVYLGALLYANVLMRSRFSDAFVARWTP
 PSAPPMPPTLPPMLPSPL
 >seq-6070
 MHALFFYHLVASVSVHVTSTLPMQLQADGEQTYQLVVPYAKHSGTDGISTSYLYKEVFGFTLSSLLIINEAITAVSHLSGLLGFLYKRQ
 MLEDDRHVEIRYYVEAITALLEVVLYVLLGGRDNALLFAIVVNTVVIQVLGYMLERTTQLRQLYLNLAGFALLLVPISFISAASLT
 DGFVSIYIYTILYVLFGLHSLLIYLSKEWRAFIDKDAGYIVLGLIAAKEILTWMVAVALQAKLYCDHGVHIKSMDYVEIDFFLQWFPIGGM
 LVAVAALFVSRFPIEDSYDTI
 >seq-6159
 VVTFDDPNSVRHFLIGFLSMFQAGCCFYMGVNRKVNSIMHTVVFICAVTACSYANWSGLGVEYKTSDETPRIVFWSRYIDWVVT
 GPLIVSALALQSKSDTPTVISLSGNVVLYEICSLIGAMTVAPYKFMWWVASLIFLGMVFMVLQMQLNNPEGYGENLQTMYLVICCF
 MYPIVWIVGSEGTAALGLSQQTGLCITDVVVKLGSFLLLNSIDEETAAPLNQNSQQYV
 >seq-6238
 MSLAIAAMGASFFFFSARGSLKEGNRMAVTLSATVVLIAFYHYVIRFDSWVATAGAGNPGAFNEGORYYVDWLLTVPLLVAELILVRL
 DKKVESKLQRNVNYYAFLMIATGYIGELDVQTSVVLDSRTIWGAISTIFFVLRLDFFGIEASFKQSKEIQQIFKVLRVVLIATWGKVPI
 AYLLPTISTSLTAGDGLVLIKQIGYTIADILAKPGYGLLWIKLTKNSGR
 >seq-6395
 MELTKLAALFLLSALAHLGNGFWVTKFYFTYLDKAQSPTRWEYFFSASVMMMSIAYTSGIRGFTELHCFWLVAITLFGWLQEVINRP
 KPDKDWEQMNMLVSRSQAHVLYIPQLAAWYSVLYTFLASASGGCGGPDFVYYIVLGEAVLFFSGAPQLYQILSPPSKYVYGEYMYQL
 LSLLSKVGLVILLVNLVYMDRFEDAVIDDNTTAEERTNCTQ
 >seq-6433
 MIGLSNLAALALAQRASVRSRSQLRWEFVSIATAVLLSVAITAIGLFFRRRTDLTLIDFGLCGGLYAIRLLTQQPTFQSLFDESRAFWH
 YADWFITCTIILPESLFLYQVNVNKNVKTFFWLFVQVQTIIFGIVAAAFGADLDKLSIANNMVVLGLTAAAIFLALFRRRTGPGESSPR
 KPWIIERIVFLAGFLIWLATVHVANLGLKVLGDHNVEFLGLFVFSLLG
 >seq-6499
 MELTLVKNTFYFTYIIFTTGTICFIEALRNSNPKIRHMNLETISVAGYFYSLFVDKLNKAELNEESDKKQELPFEEDTRYTDWVIST
 PLMLLVCLVLGMENKIVNFWNFMILAMNLMGLGVEVNTLSKNMANLIGPGFFAGVFGTIWNTMSGKRTFNSLIIYLLFVILW
 AFYGVLYQSDTLLTIFGYNILDLISKAFQIFFWLYLTKS
 >seq-6690
 MKNLKTNDFVGISFWIISIALAACTVFLVESPVYKPKWRLPVNVSIAVTGIAASHYYMMRWWVATQESPVYVYRVIDWVLTVPQIIEFY
 LILSVAKKIPIDLFLYKLLTASVLMIVGGYLGETGSIDRNLGFIVGMIAWLYIYEIFYQGAAEFEETDDESVKFTFDALKWIVTLGWSIYPI
 GYLLQTNSMNLLYNGFDLVNKLFGLIWYAGRKNNGSATLDIINSSLNTFKNI
 >seq-6860
 MTTVKVKOSLCKPYQINVAAVAGHAIVSGVFIGLYLPQKPNLVIPYTESFLKWNRADNLTSCKAGSRPLETNNSGYFCIEPTSDGINCDF
 STEPPTCDGLDGLWLISFHLLSFFFQALAGITDWPWPKFTIGGEEYKEYYNEMIEKYGNTPLRFVEYSASIMLICIAFLNGVTDLNIAIA
 VLTAMCQICGLVVEYMTDEQFWQWLHFTGWQFLCAYGIISHAFFKSIDAIPVNEPPQFVYIVFMLFALYACFGAVQLVELCQKTEQ
 FLCCCKTFLWIIPCCGCKCGIDTCDCCGVRCKRICKGEKDDNRCNDQCKELTYVILSAAKLVLTGIFTNVLFSTSEA
 >seq-6844
 MKYTEPLVNNNTLYFSIIIQITGLIAIPALFVEFLNKDDQILKDALILETIVQFIEASFYIYISLAATAIVKTPRLYFDWFITPTMLFSSIMFFYN
 KLKQDQDTVATTKNLEYDSIIIGMIMLNSAIMLLGYLGEKRRINKYLAVGIGSIFFLFSFYLYMDNYAKTILGTQLFYIFGVWSLYAVA
 ALQKTNLKNIMYNINLLSLSKNFYGLFIAIYIWTIKVNR
 >seq-8878
 MDVVTLTGYSALVNVLSFGTAVMLGAFVYFLTQMNSVSKEYRSGIAVSIAVVGIAAYHYRIWQGFSVGEMNEGORYADWLTVPL
 LVIELLIVLGVTAERKLMWRLVPATVLMIALGYPGEVSTVAETKWFALAMVPFAYILWVLYQQLNAAKKRESGAVAGAIKSATY
 >seq-8988
 MIFTKELVERTIWLSQLMQFITSIPLHAFITSKKDKLSDILALETIVQLVEMIFYIWIAYAILNINKMTSRRYIDWWVITTPIMLLSTMFMK
 YKEHITLNKEIITKKFLSDNKRQIFSIFSNLLMLIFGYFGEINILSKYIAIPLGFVFFKSFEIYKDYAIKTLQGKNLFLFLIVWSLYGVAAT
 FSPNLKVNVSYNNLLDLSVSKNFYGLIYYEILKLN
 >seq-9300
 MSIFLQVGDLTGFSSFLASMALLAATIFFLERRSVHRKGWLSTIATLITGIAALHYHMKHAWLSSSTSPTEIRYIDWLLTVPLMCIEFYL
 ILRALGHASRHILYRLFGYSIGMILLFGYLGETLIEPMLGFGVGLIFWGLIIFEVFKGEAKAMELSSHMLKSSFKLLRLFILIGWAIYPLG
 YYFGVSGGTELLNVAYNLADVINIGFSLVYLLARADSQIDSQITT