PROTEIN AS A SECOND LANGUAGE FOR LLMS

Anonymous authors

 Paper under double-blind review

ABSTRACT

Deciphering the function of unseen protein sequences is a fundamental challenge with broad scientific impact, yet most existing methods depend on task-specific adapters or large-scale supervised fine-tuning. We introduce the "*Protein-as-Second-Language*" framework, which reformulates amino-acid sequences as sentences in a novel symbolic language that large language models can interpret through contextual exemplars. Our approach adaptively constructs sequence–question–answer triples that reveal functional cues in a zero-shot setting, without any further training. To support this process we curate a bilingual corpus of 79,926 protein–QA instances spanning attribute prediction, descriptive understanding, and extended reasoning. Empirically, our method delivers consistent gains across diverse open-source LLMs and GPT-40, achieving up to 17.2% ROUGE-L improvement (average +7%) and even surpassing fine-tuned protein-specific language models. These results highlight that generic LLMs, when guided with protein-as-language cues, can outperform domain-specialized models, offering a scalable pathway for protein understanding in foundation models.

1 Introduction

Proteins are indispensable molecular machines of life, driving key functions such as maintaining cell structure and enabling cell communication. Their three-dimensional architectures, catalytic activities, interaction networks, and evolutionary trajectories are all encoded within a linear sequence composed of twenty amino-acid characters (22, 55). Therefore, the core of understanding protein function lies in accurately "reading" and "translating" the biological meaning contained within these amino-acid sequences (9, 23). However, this task is fraught with challenges. Although the amino acid sequence is formally like a language—possessing a fixed character set (over 20 genetically encoded amino acids) and potential grammatical rules (physicochemical laws)—the mapping relationship from the one-dimensional sequence to the three-dimensional structure and function is extremely complex and highly context-dependent (39, 49). Consequently, the central challenge of "what cellular function does an unknown amino acid sequence encode?" still lacks a comprehensive solution.

To address this challenge, research efforts on protein understanding can be broadly categorized into two dominant paradigms: protein representation learning and protein—language alignment modeling. Protein representation learning sees amino-acid sequences as a standalone modality like language and visual, acquires universal protein representations through self-supervised pre-training on large-scale amino-acid sequences, and then attaches lightweight decoders to predict structure or function (59, 4, 25, 42, 7, 52, 51). While this paradigm excels in the universality of its embeddings and in mining deep sequential patterns, these embeddings still rely on additional "interpreters", i.e., post-processing adapters, to be converted into human-understandable explanations. Protein—language alignment modeling, in contrast, co-trains on paired protein sequences and their textual descriptions, establishing a bidirectional mapping within a shared latent space that enables end-to-end text-based question answering (56, 37, 16, 1, 48, 54). Although this route bypasses downstream adapters, it is intrinsically bound to large-scale paired data and often requires re-fine-tuning whenever the output format or downstream objective shifts. In summary, both of these approaches face bottlenecks of large training data requirements, high computational costs, and limited generalization ability.

Protein as Second Language. Reflecting on the human cognitive process, we observe that humans exhibit remarkable efficiency and generalization ability when learning a brand-new symbolic

system (i.e., a new language). The key lies in their ability to rely on and transfer their existing native language knowledge system (14, 20). Given the aforementioned "linguistic" properties of protein sequences—possessing a compositional structure and contextual semantics—and our goal of understanding their function using natural language, we propose a novel perspective: to treat protein sequences as a symbolic system that can be learned and interpreted by large language models (LLMs) as a "second language".

Analogous to how humans acquire a second language, *i.e.*, by encountering new words in context and inferring their meaning and usage, we propose a protein language learning framework in which an LLM acquires protein semantics and reasoning ability through context-driven exposure that grounds sequence patterns in functional and structural examples. This framework adaptively constructs learning contexts for a given protein understanding goal, enabling rapid acquisition of target protein knowledge without additional training or sacrificing generalization. To support effective learning, we constructed a "bilingual" dataset of 79,926 protein-sequence—question—answer triples covering functional, descriptive, and extended-information queries. Across Protein2Text (58), Mol-Instructions(12) and ProtDescribe-QA (2), our framework raises the average ROUGE-L by 7% across diverse open-source models and GPT-40, with a maximum gain of 17.2%, without any task-specific fine-tuning. Our contributions are as follows:

- We introduce the "*Protein-as-Second-Language*" conceptual framework, which recasts aminoacid sequences as a second language that can be acquired via in-context learning, enabling efficient and generalized protein understanding.
- We construct *a protein-natural language bilingual dataset* that spans four task families: attribute-based QA, True or False QA, descriptive-text QA, and extended-information QA, to support effective protein language learning and benchmarking.
- We present a protein language learning framework that adaptively constructs learning contexts
 for protein understanding, yielding significant gains for both open-source models and GPT-4o,
 enabling them to outperform domain-specialized models without additional training.

2 RELATED WORK

2.1 Language Models in Protein

Protein representation learning with protein language models (PLMs) extends the Transformer to amino-acid strings, producing dense embeddings for property prediction (17, 4, 11, 18, 5, 7, 8) or generative design (32, 34, 29, 13). Because these models are trained exclusively on amino acid sequences, their outputs remain latent vectors that external classifiers must translate into human-readable function. To obviate this indirection, protein—language alignment modeling has emerged, which jointly connects sequences with textual descriptions via (i) contrastive objectives mapping proteins and sentences into a shared space (57, 51), (ii) bioknowledge-augmented pre-training on curated protein—text corpora (13, 44, 29, 36, 60, 26), or (iii) multi-modal LLMs that graft protein encoders onto frozen language backbones (27, 1, 48, 8, 31, 53). While effective, these approaches entail costly retraining or gradient updates and risk catastrophic forgetting when scaled to larger LLMs (21, 50), prompting a shift toward parameter-efficient adaptation.

2.2 PROTEIN QA DATESETS

Datasets that couple proteins with natural-language annotations have become the empirical bedrock for developing protein–text hybrid systems. At present, two complementary families of corpora dominate the landscape. The first centers on protein captioning: given an amino-acid sequence alone, the objective is to generate a concise textual description. Representative instances include the richly annotated Swiss-Prot collection (3), the ProteinKG resource (59) and ProtDescribe (58). The second family targets protein question answering: here, both a sequence and a natural-language query are supplied, and the model is required to synthesize an answer grounded in the provided protein. Curated examples span Mol-Instructions (12), UniProtQA (28), ProteinLMBench (40), VenusX (43) and Protein2Text-QA (2).

3 PROTEIN AS SECOND LANGUAGE

We introduce "Protein-as-Second-Language", a framework that treats amino-acid sequences as a new symbolic system to be learned much like humans acquire a foreign language. Just as learners infer the meaning of unfamiliar words by repeatedly encountering them in context, we construct a *protein-natural language bilingual dataset* (Sec. 3.1) and design an *adaptive context construction mechanism* (Sec. 3.2) to provide such contextual exposure. In this way, our framework enables LLMs to acquire protein semantics through exemplars rather than through extensive re-training.

3.1 BILINGUAL DATASET CONSTRUCTION

We curate our bilingual dataset in three steps (Figure 1). Starting from 573,661 Swiss-Prot (3) entries with gene ontology (GO) annotations, we avoid directly converting all annotations, as this would introduce heavy redundancy; instead, we construct a balanced sample. Specifically, (i) we prune the GO-directed acyclic graph (GO-DAG) to obtain representative functional categories and group proteins accordingly (Sec. 3.1.1), (ii) perform bilingual deduplication by clustering sequences within each protein group and sampling proteins with diverse functional annotation (Sec. 3.1.2), and (iii) use DeepSeek-R1 (15) to generate attribute, knowledge, descriptive, and true/false QA pairs, yielding 79,926 high-quality protein–QA triples (Sec. 3.1.3).

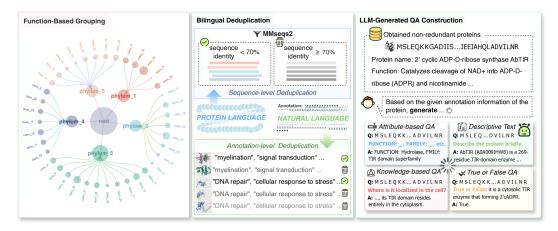


Figure 1: The overview of data construction of our bilingual protein-QA dataset.

3.1.1 Function-Based Grouping

To enable representative sampling across functional categories, the dataset is partitioned according to the GO hierarchy. Directly using the raw directed acyclic graph (DAG) risks over-fragmentation from overly fine, sparsely populated terms, and excessive generalization near the root. To address this, we adapt a pruning strategy inspired by decision tree simplification (33), where complexity is managed through a penalty to avoid overfitting. This strategy aims to retain an optimal set of GO terms as functional grouping nodes. It balances granularity and coverage, ensuring that the retained nodes represent biologically diverse yet statistically well-supported categories for downstream sampling.

The pruning process is driven by two main criteria: (1) A node is retained if it meets the *minimum support threshold*, which ensures that the node has a sufficient number of associated proteins, and does not exhibit significant child imbalance. (2) If the *child-imbalance ratio* is high, meaning the protein distribution among a node's child terms is uneven, the parent node is retained, even if the child nodes fail to meet the minimum support threshold.

Minimum Support Threshold A node is retained only if the number of associated proteins meets a depth-adjusted threshold m(d), which adapts based on the node's depth in the GO hierarchy. The threshold is calculated as:

$$m(d) = \lambda \cdot C_{tot} \cdot (1 + \beta d) \tag{1}$$

where C_{tot} is the total protein count, d is the node depth, and λ and β are constants. This dynamic threshold is designed to prevents deep nodes from splitting infinitely due to overly small absolute values.

Child-Imbalance Ratio The child-imbalance ratio is applied to assess whether the child nodes of a given term are too imbalanced. The imbalance ratio $\rho(v)$ is computed as the ratio of the largest to the smallest protein count among the child nodes:

$$p(u) = \frac{\max_{u \in C^{+}(v)} C(u)}{\min_{u \in C^{+}(v)} C(u)}$$
 (2)

where $C^+(v)$ represents the set of valid child nodes with non-zero protein counts. If the imbalance ratio $\rho(v)$ exceeds a specified threshold $\tau(d)$, the parent node v is retained to preserve the biological diversity. This threshold is adjusted dynamically with the depth d to allow for greater flexibility at deeper levels of the hierarchy:

$$\tau(d) = \tau_0 \cdot \alpha^d \tag{3}$$

where τ_0 is the base threshold, and α is a scaling factor.

By applying these two criteria, the pruning process is carried out recursively, allowing the algorithm to adaptively prune the GO DAG and identify the most relevant, biologically diverse functional groups.

3.1.2 BILINGUAL DEDUPLICATION

After grouping by GO term, proteins within the same node often exhibit high similarity, as they represent homologous proteins. To address this, we use MMseqs2 (41) for sequence clustering within each GO node, applying a 70% *amino acid sequence similarity* threshold. From each cluster, a single representative sequence is selected. This threshold efficiently removes redundant sequences with minimal functional variation while preserving functional diversity.

While sequence similarity-based redundancy removal effectively reduces sequence-level redundancy, it does not necessarily capture functional divergence. Specifically, sequence similarity below 70% does not imply functional divergence, and substantial functional redundancy may still exist within the set (10). To address this, we focus on *annotation semantic similarity*, quantifying the functional relationships between proteins based on their GO annotations. Inspired by the simGIC method (38) for calculating GO terms semantic similarity, we calculate the Protein Functional Information Content IC protein function for each protein, which is the sum of the Information Content (IC) of all associated GO terms and their ancestral terms. The IC of each GO term is calculated based on its frequency in the dataset, using the total protein set after sequence redundancy removal. The IC protein function value for each Protein ID is computed as:

$$IC_{\text{protein function}} = \sum_{g \in GO \text{ terms of } p} IC(g) + \sum_{g' \in \text{ancestors of } GO \text{ terms of } p} IC(g'). \tag{4}$$

This provides a quantitative measure of each protein's functional information, capturing both direct and indirect annotations. For each GO term, proteins are sampled based on their unique IC_{protein function} values (rounded to 3 decimal places). To ensure balanced species representation, a species quota strategy is applied based on the proportions of Eukaryota, Bacteria, Archaea, and Viruses in the dataset after sequence redundancy removal. This ensures an unbiased species distribution in the final sample. The bilingual deduplication process reduces redundancy in two aspects, amino acid sequence and annotation semantics, ensuring a balanced and diverse protein corpus.

3.1.3 LLM-BASED QA CONSTRUCTION

To transform curated protein annotations into natural-language question—answer pairs, we prompt the DeepSeek-R1 (15) model to generate biologically grounded QA texts that reflect both functional attributes and contextual knowledge (the prompts used for each QA type are provided in

Appendix D). The resulting QA corpus covers four complementary types: ① Attribute-based QA captures factual properties directly associated with a protein, such as molecular function, cellular component, or family. ② Knowledge-based QA comprises concise, annotation-driven questions and answers that involve in multiple biological aspects of a protein, such as expression, localization, mechanism, and interactions. ③ Descriptive Text QA produces longer natural-language explanations that integrate multiple annotations into coherent functional summaries. ④ True or False QA consists of single statements that integrate multiple biological aspects of a protein, accompanied by a True/False answer and a brief explanation.

These four types yield a rich and varied bilingual dataset, ensuring that models are exposed to both concise factual knowledge and more detailed contextual explanations, supporting their ability to understand and reason about protein functions.

3.2 BILINGUAL CONTEXTUAL LEARNING

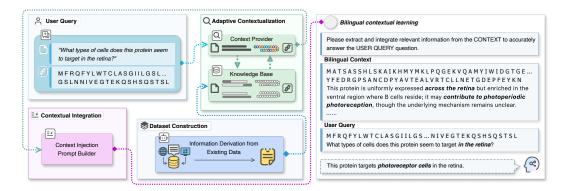


Figure 2: Process of Query-Adaptive Context Construction.

In practical scenarios, questions concerning protein sequences are often highly flexible and complex: they require not only analogous proteins with similar sequence patterns to capture potential structural or functional signals, but also complementary descriptive knowledge and QA pairs to provide semantic grounding. As shown in Figure 2, we propose an adaptive context construction mechanism, for *bilingual contextual learning*, designed to selectively build bilingual learning contexts for each query. Instead of brute-force mixing of amino acid sequences and descriptive texts, the mechanism follows the principle of second language acquisition—exposing learners to new words in context so that meaning and usage can be inferred (19). By analogy, LLMs acquire protein semantics and reasoning ability through context-driven exposure that grounds sequence patterns in functional and structural exemplars.

The mechanism operates in three stages. First, the adaptive context provider selects candidate contexts from the protein–natural language bilingual dataset, guided by two criteria: (i) amino acid sequence homology between candidate proteins and the query sequence, computed with MM-seqs2 (41), and (ii) similarity between the descriptive texts or QA pairs of candidate proteins and the query question. Second, the contextual integration module structures the selected examples into a coherent context. Finally, the constructed bilingual context is combined with the query and presented to the LLM as in-context examples, enabling analogy-based reasoning and evidence integration to produce biologically meaningful responses.

4 EXPERIMENTS

4.1 SETUP

Evaluation Datasets We comprehensively evaluated our method using 3 text-based protein understanding datasets: • ProtDescribe (58) comprises 553,052 high-quality protein—text pairs extracted from Swiss-Prot. Each instance pairs an amino-acid sequence with a single textual description obtained by concatenating four annotation fields in a fixed order: protein name, function, subcellular

location, and similarity. The resulting descriptions average 40–60 tokens. Protein2Text-QA (2) comprises 209,847 open-ended question—answer pairs covering 5,574 unique proteins. Each instance consists of an amino-acid sequence, a free-form question, and a concise answer; all QAs are automatically generated from PubMed abstracts/discussion/introduction sections and presented as conversational natural-language text without fixed templates. Mol-Instructions (12) comprises 2.04 M instruction instances divided into three major sections: molecule-oriented, protein-oriented, and biomolecular-text. The protein-oriented section alone contributes 505 K instructions covering diverse tasks. Each sample is formatted as a natural-language "instruction—input—output" triplet: the input is a UniProt amino-acid sequence, and the output is a free-text answer tailored to the specific task.

Models All experiments are conducted under identical prompting protocols. We first evaluate the proposed adaptive context construction method on frozen LLMs, including Qwen2.5-3B (46), Mistral-7B-Instruct-v0.3 (6), Qwen3-14B (47), Kimi-k2 (45), and GPT-4o (35), to test few-shot and compositional reasoning capabilities, thereby mimicking the dynamics of second language acquisition. In addition, we also evaluate fine-tuned protein-oriented LLMs, such as BioT5-plus-base (37) and ProLLaMA (30), which have been explicitly trained on large-scale protein corpora. These models serve as a baseline for comparison, allowing us to examine the performance gains of our method in general-purpose frozen LLMs relative to specialized protein LLMs.

Metrics We evaluate model outputs using both an automatic metric (ROUGE-L (24)) and human evaluation. ROUGE-L (24), though widely used for text generation, primarily measures lexical overlap and may not fully capture semantic correctness in protein-related QA. To address this limitation, five evaluators rated the quality of generated answers on a 0–5 scale, where 0 denotes garbled and unreadable content, intermediate scores reflect increasing levels of informativeness and accuracy, and 5 represents fully correct outputs (detailed scoring rubrics are provided in Appendix A). This combined evaluation provides a more reliable assessment of factual accuracy and overall comprehensibility.

4.2 QUALITY OF DATASET

Figure 3 (a-f) provides a multidimensional analysis of the protein sequences included in our dataset. The collection spans a wide range of sequence lengths, from short peptides to large multi-domain proteins, and covers proteins from 4,135 species across diverse evolutionary lineages. At the family level, the dataset comprises 63,749 families and 1,115 superfamilies, ensuring representation of both well-studied proteins and rare functional groups. Additional annotations capture domain composition, catalytic activity classes, and gene ontology categories, collectively highlighting the long-tail distribution across sequence space and functional categories. This diversity ensures broad biological coverage while posing realistic challenges in inferring functions for proteins, particularly for infrequent families and underexplored functions.

Figure 3 (g,h) summarizes the distribution of tasks and token composition within the dataset. The corpus encompasses four distinct protein-QA types, with sample counts ranging from 11,693 (attribute-based QA) to 32,444 (true/false QA), thereby providing balanced coverage across multiple functional perspectives. In terms of token composition, amino-acid sequences constitute nearly 70 % of the corpus, reflecting the sequence-centric nature of protein understanding tasks and highlighting the need for models to align symbolic sequence information with natural-language context effectively.

4.3 MAIN RESULTS

Accuracy gains from context-driven exposure Table 1 presents that our method consistently improves performance on three text-based protein understanding datasets. Our method raises the average ROUGE-L by 7% across diverse open-source models and GPT-40 (35), with a maximum gain of 17.2%, demonstrating that context-driven exposure allows LLMs to acquire protein semantics and reason about function directly from sequence and textual context without any parameter updates. Larger models benefit more, suggesting that greater capacity enhances the ability to leverage contextual cues, consistent with learning protein meaning through in-context analogy and reasoning. In contrast, fine-tuned protein LLMs such as ProLLaMA-7B (30) do not surpass frozen

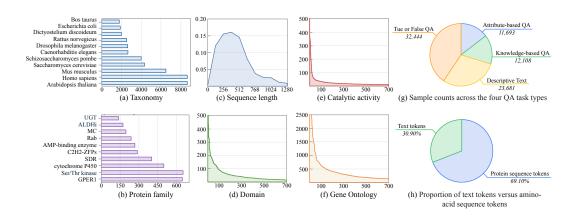


Figure 3: **Dataset statistics.** Left: Multidimensional analysis of protein amino-acid sequences, including length, domain composition, and catalytic activity. Right: Sample sizes for the four protein-QA types and the ratio of textual to amino-acid sequence tokens.

Table 1: Comparison of different approaches in protein question answering "Fun.", "Des.", "Dom.", and "Cat." denote the 4 protein-oriented tasks in the Mol-Instructions dataset (12): protein function prediction (Fun.), general textual description generation (Des.), domain/motif recognition (Dom.), and catalytic activity prediction (Cat.). Δ *Gain* shows the percentage performance increase. \Diamond indicates LLMs augmented with our adaptive context construction method. Metric: ROUGE-L.

Model	ProtDescribe	Protein2Text-	Mol-Instructions					
	ProtDescribe	QA	Func.	Desc.	Dom.	Cat.	Avg.	
Fine-tuned LLM								
BioT5+ (37)	9.97	6.96	2.92	6.22	2.37	2.87	3.60	
ProLLaMA-7B (30)	12.77	10.09	16.89	15.34	15.85	19.32	16.85	
Frozen LLM								
Qwen2.5-3B (46)	18.45	23.21	18.91	17.18	18.01	20.05	18.54	
Qwen2.5-3B (46) \diamondsuit	27.32	28.66	22.05	22.23	25.14	15.96	21.35	
Δ Gain	+8.87	+5.45					+2.81	
Mistral-7B-Instruct-v0.3 (6)	15.02	20.97	17.05	18.59	14.95	18.07	17.17	
Mistral-7B-Instruct-v0.3 (6) ♦	29.39	28.59	15.77	22.72	17.46	21.20	19.29	
Δ Gain	+14.37	+7.62					+2.12	
Qwen3-14B (47)	23.20	21.02	15.80	12.75	15.81	14.06	14.61	
Qwen3-14B (47) \diamondsuit	35.53	25.93	20.17	17.37	18.47	23.25	19.82	
Δ Gain	+12.33	+4.91					+5.21	
kimi-k2 (45)	26.74	17.33	12.60	12.36	10.32	15.97	12.81	
kimi-k2 (45) ♦	35.91	21.04	14.47	14.97	15.68	17.02	15.54	
Δ Gain	+9.17	+3.71					+2.72	
GPT-4o (35)	18.29	20.84	16.89	14.50	16.74	20.00	17.03	
GPT-4o (35) ♦	35.53	26.86	20.24	19.23	17.46	22.61	19.89	
Δ Gain	+17.22	+6.02					+2.85	

LLMs augmented with our method, likely due to limited training coverage and task-specific rigidity. This underscores the our method as a lightweight alternative that enables general-purpose LLMs to potentially exceed the performance of domain-adapted models.

Human evaluation further demonstrates that exposing models to curated protein–language contexts improves the perceived quality of outputs (Figure 4). Across all rated instances, inter-rater consistency was substantial (Krippendorff's $\alpha = 0.72\%$), ensuring reliable annotations; the detailed rubric is given in Appendix A. Models receiving context-driven exposures achieve higher or comparable

ratings on most tasks (left panel), with the clearest improvements observed on Protein2Text-QA (2) and several Mol-Instructions (12) subtasks. Furthermore, pairwise win/lose analyses (right panel) show that outputs generated with context-driven exposure are preferred in the majority of comparisons, with win rates systematically exceeding loss rates.

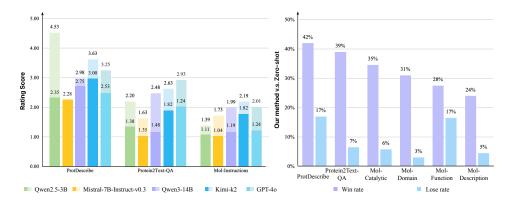


Figure 4: **Comparison of human evaluation results.** Left: Absolute human rating scores (0–5) for zero-shot model outputs (dark bars) and model outputs with adaptive context exposure (light bars) on three datasets. Right: Pairwise win/lose proportions comparing outputs with and without adaptive context exposure. Each comparison is based on 8 randomly selected cases per subset (48 cases in total across six subsets).

Varying exemplar number (k) Figure 5 illustrates how model performance varies with the number of exemplars (k) provided in context across different datasets. Performance generally improves as k increases, but only up to a task-dependent optimum; beyond this point, additional exemplars offer little benefit or even introduce noise. The optimal k differs by task. For ProtDescribe (58), which involves fixed attribute-centric questions, a larger set of bilingual exemplars from related proteins helps the model capture recurring patterns, with performance peaking at k=10–11. In contrast, Protein2Text-QA (2) requires open-ended and integrative reasoning, where only a small number of highly relevant exemplars are beneficial; here, performance peaks earlier at k=3–4. In our experiments, we therefore adopt the task-specific optimal settings: k=11 for ProtDescribe (58), k=4 for Protein2Text-QA (2), and k=4 for Mol-Instructions (12).

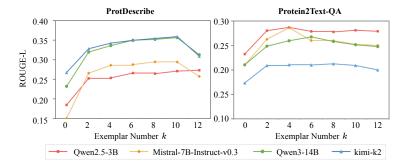


Figure 5: Effect of varying exemplar number (k) on model performance. We explored $k \in [1, 12]$ as the search space; the upper bound was set after a coarse scan up to k = 50 showed performance saturation around 2-12 exemplars. Metric: ROUGE-L.

Ablation on dual-criterion context selection Table 2 shows that using both sequence homology and text/QA similarity (Dual) outperforms either criterion alone, providing complementary signals that maximize the effectiveness of context-driven exemplar selection. On average across three datasets, using only sequence homology reduces performance by 5.2%, and using only text/QA similarity reduces performance by 2.8% compared to Dual, though all variants still outperform zero-shot models.

Table 2: **Ablation on dual-criterion context selection.** Columns show model performance when using both sequence homology and text/QA similarity (Dual), sequence homology only (SeqOnly), or text/QA similarity only (QAOnly) to guide exemplar selection. Metric: ROUGE-L.

Model	ProtDescribe		Protein2Text-QA			Mol-Instructions			
	Dual	SeqOnly	QAOnly	Dual	SeqOnly	QAOnly	Dual	SeqOnly	QAOnly
Qwen2.5-3B (46)	27.32	20.10	25.24	28.66	27.76	25.73	21.35	20.16	19.68
Mistral-7B-Instruct-v0.3 (6)	29.39	19.46	25.24	28.59	21.85	22.68	19.12	14.31	17.40
Qwen3-14B (47)	35.53	22.90	30.52	25.93	23.26	25.87	19.82	15.07	17.58
kimi-k2 (45)	35.91	28.58	32.43	21.04	17.10	18.06	15.54	12.79	12.09
GPT-4o (35)	35.51	27.72	32.59	26.86	23.90	26.45	19.89	17.46	17.08

Case studies and qualitative evaluation Figure 6 illustrates that context-driven exposure produces concise, function-specific descriptions consistent with UniProt annotations. In the two examples shown, the model correctly identifies "intrinsically disordered regions", and "[4Fe-4S] RNA methyltransferase activity", whereas zero-shot outputs remain generic.

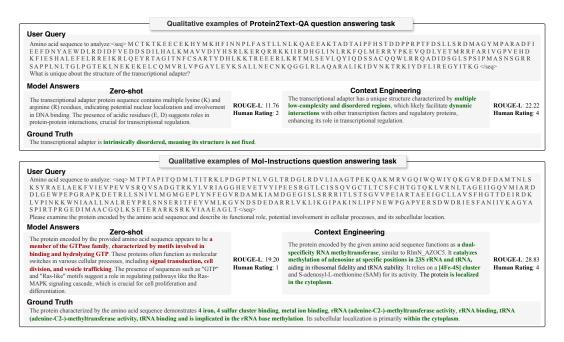


Figure 6: **Qualitative examples of protein question answering.** We present two examples with answers generated by GPT-4o (35) along with the target ground truth. The green color highlights accurate keywords, while the red color indicates prediction errors.

5 CONCLUSION

We have proposed the "Protein-as-Second-Language" framework, which leverages adaptive context construction to enhance bilingual protein understanding by dynamically integrating sequence homology and textual similarity. Additionally, we introduced a protein-natural language bilingual dataset, specifically designed to support this framework and facilitate the bridging of protein sequences with functional descriptions. Our approach has successfully enhanced large language models' ability to acquire protein semantics and reasoning capabilities without the need for task-specific parameter updates. Experiments on multiple protein-language datasets demonstrate that our framework consistently outperforms zero-shot baselines, highlighting the effectiveness of context-driven learning in bridging protein sequences with functional descriptions.

6 ETHICS STATEMENT

This work complies with ethical standards and established research practices. All protein data were sourced from publicly available databases, with no proprietary or confidential information involved. Quality assurance and safety checks were applied to minimize harmful or inappropriate content. We acknowledge the broader risks of combining LLMs with biomolecular knowledge, including potential misuse for harmful purposes, and therefore emphasize responsible use guided by fairness, transparency, and accountability. Any harmful or unsafe applications of this dataset are strictly prohibited.

7 REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the protein–natural language bilingual dataset (Sec. 3.1, Appendix C), the adaptive context construction mechanism (Sec. 3.2). Data processing steps and QA generation prompts for all four question types are included in Sec. 3.1 and Appendix D. Code implementing the framework and instructions for reproducing experiments on both frozen and protein-adapted LLMs will be provided as supplementary material upon acceptance.

REFERENCES

- [1] Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10757–10765, 2024.
- [2] David Arredondo Yue Hu Luis Tafoya Kushal Virupakshappa Avinash Sahu Ala Jararweh, Oladimeji Macaulay. Protein2text: Resampling mechanism to translate protein sequences into human-interpretable text. In NAACL 2025 - Industry Track, 2025.
- [3] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.
- [4] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.
- [5] Yue Cao and Yang Shen. Tale: Transformer-based protein function annotation with joint sequence–label embedding. *Bioinformatics*, 37(18):2825–2833, 2021.
- [6] Devendra Singh Chaplot. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3, 2023.
- [7] Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.
- [8] Zhiyuan Chen, Tianhao Chen, Chenggang Xie, Yang Xue, Xiaonan Zhang, Jingbo Zhou, and Xiaomin Fang. Unifying sequences, structures, and descriptions for any-to-any protein generation with the large multimodal model helixprotx. *arXiv* preprint arXiv:2407.09274, 2024.
- [9] Wyatt T Clark and Predrag Radivojac. Analysis of protein function and its prediction from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 79(7):2086–2096, 2011.
- [10] Damien Devos and Alfonso Valencia. Practical limits of function prediction. *Proteins: Structure, Function, and Bioinformatics*, 41(1):98–107, 2000.

- [11] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
 - [12] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/pdf?id=Tlsdsb619n.
 - [13] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.
 - [14] Susan M Gass, Jennifer Behney, and Luke Plonsky. Second language acquisition: An introductory course. Routledge, 2020.
 - [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
 - [16] Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. *Authorea Preprints*, 2023.
 - [17] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, 2025.
 - [18] Bozhen Hu, Cheng Tan, Yongjie Xu, Zhangyang Gao, Jun Xia, Lirong Wu, and Stan Z. Li. Protgo: Function-guided protein modeling for unified representation learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 88581–88604. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a1722a6bd1023c026a3d6a570fb3af75-Paper-Conference.pdf.
 - [19] Thomas Huckin and James Coady. Incidental vocabulary acquisition in a second language: A review. *Studies in second language acquisition*, 21(2):181–193, 1999.
 - [20] Scott Jarvis and Aneta Pavlenko. *Crosslinguistic influence in language and cognition*. Routledge, 2008.
 - [21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
 - [22] Norio Kitadai and Shigenori Maruyama. Origins of building blocks of life: A review. *Geoscience Frontiers*, 9(4):1117–1153, 2018.
 - [23] Eugene Koonin and Michael Y Galperin. Sequence—evolution—function: computational approaches in comparative genomics. 2002.
 - [24] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 workshop on automatic summarization*, pp. 45–51, 2002.
 - [25] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.
 - [26] Nuowei Liu, Changzhi Sun, Tao Ji, Junfeng Tian, Jianxin Tang, Yuanbin Wu, and Man Lan. Evollama: Enhancing llms' understanding of proteins via multimodal structure and sequence representations. *arXiv preprint arXiv:2412.11618*, 2024.

- [27] Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. ProtT3: Protein-to-text generation for text-based protein understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5949–5966, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long. 324. URL https://aclanthology.org/2024.acl-long.324/.
- [28] Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [29] Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. Prollama: A protein large language model for multi-task protein language processing. *arXiv e-prints*, pp. arXiv–2402, 2024.
- [30] Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. Prollama: A protein large language model for multi-task protein language processing. *IEEE Transactions on Artificial Intelligence*, 2025.
- [31] Zicheng Ma, Chuanliu Fan, Zhicong Wang, Zhenyu Chen, Xiaohan Lin, Yanheng Li, Shihao Feng, Jun Zhang, Ziqiang Cao, and Yi Qin Gao. Prottex: Structure-in-context reasoning and editing of proteins with large language models. *arXiv preprint arXiv:2503.08179*, 2025.
- [32] Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature echnology*, 41(8):1099–1106, 2023.
- [33] Camille Mondon. Classification and regression trees. 1984.
- [34] Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.
- [35] OpenAI et al. Gpt-4 technical report, 2024.
- [36] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 1102–1123, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.70. URL https://aclanthology.org/2023.emnlp-main.70/.
- [37] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv* preprint arXiv:2310.07276, 2023.
- [38] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(Suppl 5):S4, 2008.
- [39] Burkhard Rost et al. Protein structure prediction in 1d, 2d, and 3d. *Encyclopedia of Computational Chemistry*, pp. 2242–2255, 1998.
- [40] Yiqing Shen, Zan Chen, Michail Mamalakis, Luhan He, Haiyang Xia, Tianbin Li, Yanzhou Su, Junjun He, and Yu Guang Wang. A fine-tuning dataset and benchmark for large language models for protein understanding. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2390–2395. IEEE, 2024.
- [41] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.
- [42] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pp. 2023–10, 2023.

- Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. Venusx: Unlocking fine-grained functional understanding of proteins, 2025. URL https://arxiv.org/abs/2505.11812.
 - [44] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv preprint*, abs/2211.09085, 2022. URL https://arxiv.org/abs/2211.09085.
 - [45] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv* preprint arXiv:2507.20534, 2025.
 - [46] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.
 - [47] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505. 09388.
 - [48] Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. Protchatgpt: Towards understanding proteins with large language models. *arXiv preprint arXiv:2402.09649*, 2024.
 - [49] Zhengyuan O Wang and David D Pollock. Context dependence and coevolution among amino acid residues in proteins. In *Methods in enzymology*, volume 395, pp. 779–790. Elsevier, 2005.
 - [50] Juntong Wu, Zijing Liu, He Cao, Hao Li, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. Rethinking text-based protein understanding: Retrieval or llm? arXiv preprint arXiv:2505.20354, 2025.
 - [51] Kevin E Wu, Howard Chang, and James Zou. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pp. 2024–05, 2024.
 - [52] Kevin E Wu, Kathryn Yost, Bence Daniel, Julia Belk, Yu Xia, Takeshi Egawa, Ansuman Satpathy, Howard Chang, and James Zou. Tcr-bert: learning the grammar of t-cell receptors for flexible antigen-binding analyses. In *Machine Learning in Computational Biology*, pp. 194– 229. PMLR, 2024.
 - [53] Wenkai Xiang, Zhaoping Xiong, Huan Chen, Jiacheng Xiong, Wei Zhang, Zunyun Fu, Mingyue Zheng, Bing Liu, and Qian Shi. Fapm: functional annotation of proteins using multimodal models beyond structural modeling. *Bioinformatics*, 40(12):btae680, 2024.
 - [54] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal Ilm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*, 2024.
 - [55] Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, et al. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.
 - [56] Hanwen Xu and Sheng Wang. Protranslator: zero-shot protein function prediction using textual description. In *International conference on research in computational molecular biology*, pp. 279–294. Springer, 2022.
 - [57] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38749–38767. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23t.html.
 - [58] Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

- [59] Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.
- [60] Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8950–8963, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.484. URL https://aclanthology.org/2024.acl-long.484/.

A EVALUATION METRICS

We use the automatic metric ROUGE-L (24) to assess the quality of the generated text by comparing it with reference answers. In addition, we incorporate manual checking into the evaluation pipeline and compute a human-rating score. Five evaluators with biological-research experience were asked to rate each generated answer on a 0–5 scale (the integer score corresponds to the category number minus one). All evaluators have at least two years of research experience in biology. The six ordinal categories they used are:

- 1. **Garbled** the text is incomprehensible and lacks any readability.
- 2. **Inaccurate** the text is readable but entirely incorrect and devoid of meaningful information.
- 3. **Partially informative** the text offers some reference value, yet its factual correctness is poor.
- 4. **Moderately accurate** roughly half of the information is correct, but several errors remain.
- Mostly accurate the content is almost entirely correct, with only minor omissions or errors.
- 6. **Completely correct** the content is accurate in its entirety, without any mistakes.

B ADDITIONAL RESULTS

Evaluation on real-world protein scenarios To examine the applicability of our framework beyond benchmark datasets, we evaluated it on biologically relevant queries involving uncharacterized *Homo sapiens* proteins. For each case, a current biologically relevant question of research interest was paired with the corresponding protein amino acid sequence and input to representative LLMs guided by our framework. As illustrated in Figure 7, the models produced plausible hypotheses aligned with biological knowledge. These results demonstrate that our framework can extend to real-world scenarios, offering interpretable preliminary insights into proteins lacking experimental annotation and potentially guiding future biological investigations.

C BILINGUAL DATASET: DETAILS AND EXAMPLES

All protein entries are derived from the manually curated SwissProt (3) section of UniProt, which provides high-quality annotations of protein sequences and functional descriptions. After deduplication, we prompted LLMs to generate four types of bilingual QA pairs from these entries. To assess data quality, we randomly sampled 500 pairs from the full set of 79,926 automatically constructed examples. Each sampled pair was manually reviewed by domain experts along three dimensions: semantic fidelity, biological plausibility, and translation fluency. The review showed a 95% pass rate, confirming that the dataset maintains high linguistic accuracy and biological reliability. Examples of four bilingual QA types are provided in Figure 8 and Figure 9.

```
Protein id: A0A1B0GW15

Open Research Question

Does it function intracellularly (as a regulatory microprotein) rather than as a secreted signaling peptide, and what cellular process might it influence?

Uncharacterized protein (LOC122394732) [Homo sapiens]: -seep-M G F V T N K S A F K A G D S L Y L R R A F V N N L G E E R R T R I Q I Q S I Q K A L D I Q I R E I D R E K A A L K R F L V K L H K T T G Y F P Q K P L W-(seep-

Model Answers

Qwen3-148: No, it does not function intracellularly as a regulatory microprotein but rather as a secreted signaling peptide. The cellular process it might influence could involve signaling pathways related to extracellular communication, such as immune responses, synaptic plasticity, or vasoconstriction, depending on its specific interactions and ligands.

Kimi-k P. No. the sequence contains a predicted signal peptide (MGFVTNKSAF) and lacks known nuclear localization or other intracellular retinon motifs. It is most likely secreted and, based on its length and basic charge, may act as an extracellular signaling peptide influencing cell-cell communication or receptor-mediated processes rather than intracellular regulation.

GP I-40: The sequence does not appear to be a secreted signaling peptide based on its length and composition. Secreted peptides are typically shorter and contain signal sequences that facilitate their export out of the cell. This protein sequence is relatively long and lacks obvious signal peptide motifs.

Protein id: PODMB2

Open Research Question

Is the protein involved in cellular stress or growth signaling pathways, adjusting protein synthesis in response to conditions?

Uncharacterized protein CS0788 [Homo sapiens]: -seep-METK K L L G K P L Q P A R P V R H L T S P G A V F P F N P Q N E Y P C N T Q C I Q S G V S R C K T N G M Q A F S Q G L N E Q Q Q Q S P V K K E I I K Y S R D F L L K L S S V S I C R K K P D F L P D H P I V L Q K P E N N Q S F K -/seep-

Model Answers

Qwen3-14B: There is no evidence to suggest that the uncharacterized protein CS0788 is
```

Figure 7: Exploratory evaluation of our framework on real-world protein scenarios.

D LLM STATEMENT

We acknowledge the use of LLMs in this work. Specifically, DeepSeek-R1 (15) was employed for two purposes: (i) polishing the English presentation of the manuscript, and (ii) generating bilingual dataset entries from curated protein annotations, where the prompts were carefully designed to ensure scientific accuracy and linguistic quality. Below we provide the exact prompts used for each bilingual QA type in the dataset construction process.

Prompt for Attribute-based Answer generation is following:

```
"Based on the provided annotations, compose a concise protein information description in the following fixed format:
PROTEIN NAME: ...

FROTEIN NAME: ...

SUBCELLULAR LOCATION: ...

FAMILY: ...

KEY SEQUENCE MOTIF: ... (write N/A if none).

After the fixed fields, leave one blank line and proceed to the 'Extended Information' paragraph. In fluent, professional English, supply any additional details essential for understanding the protein, integrating all relevant annotation content in a coherent narrative. Maintain brevity and avoid redundancy."
```

Prompt for True or False QA generation is following:

```
"You are a protein science expert. Please read the UniProt entry above and design 1 True/False question that meets all of the following rules:

(1) The stem must weave together diverse distinct knowledge dimensions from the entry (e.g., catalytic chemistry, structural biology, disease relevance, evolutionary conservation, PTM, mutational effect, regulatory mechanism, substrate selectivity, experimental evidence, GO term, PDB ID, cofactor, physiological pathway, drug-target potential).

(2) Do not include the words 'True/False' in the stem; hide the decisive technical point within the details.

(3) Give True or False, followed by an explanation.

Use this exact output template: Stem: ...; Answer: ...; Explanation: ..."
```

Prompt for Descriptive Text generation is following:

```
"Based on the given annotation information of the protein, describe the given amino-acid sequence in one coherent paragraph that covers:

(1) its main function and any bound cofactors/ligands,

(2) subcellular localization and the biological process involved,

(3) key domains or motifs, and

(4) the specific reaction catalyzed if it is an enzyme. The description begins with A sentence pattern like

'A short report on the protein with the given amino acid sequence highlights:'

or 'A brief overview of the protein with the provided amino acid sequence is as follows:'

or 'A concise description of the protein with the specified amino acid sequence includes:'

or 'An outline of the key aspects of the protein with the corresponding amino acid sequence is as follows:'

or 'A summary of the protein's main attributes with the input amino acid sequence reveals:'

(uses similar synonymous sentences to avoid uniformity)."
```

Examples of attribute-based protein-QA Protein id: O29DY1 Kinesin-like protein Klp68D; <seq>MSAKSRRPGTASSQTPNECVQVVVRCRPMSNRERSEGSPEVVNVYPNRGVVELQNVVDANKEQRKVFTYDAAYD ASASQTTLYHEVVFPLVSSVLEGFNGCIFAYGQTGTGKTFTMEGVRGNDDLMGIIPRTFEQIWLHINRTENFQFLVDVSYLEIYMEELRDLLK PNSKHLEVRERGGGVYVPNLHAINCKSVDDMIRYMKVGNKNRTVGFTNMNEHSSRSHAIFMIKIEMCDTETNTIKVGKLHIDLAGSEROSK TGASAERLKEASKINLALSSLGNVISALAESSPHVPYRDSKLTRLLQDSLGGNSKTIMIANIGPSNYNYNETLTTLTKYASRAKSIQNQFIKNEDP TGASAERLKEASKINLALSSLGNVISALAESSPHVPYRDSKLTRLLQDSLGGNSKTIMIANIGPSNYNYNETLTTLRYASRAKSIQNQPIKNEDI QDAKLKEYQEEIERLKRLIAPQQQQRSEK GGTIKKQRVKKPKKEPPISQELIGSALQASSADLQVDEDRDSDGDGAESESDKKENEAEVAKSNE ELERERVENAKLAAKLAELEGQLVRGGKNLLDTYSERQIELEKKLVEIAERKKREIEIQQQLELQEETTLEIRERNVSLEQEVELKKRKLSKC YAKYLALQQELNDCKHDHNQDLRELEMAQNELVKELKRQLLIIDNFVPIEVKQRLYTQAKYDEEQEEWKFSSFPLPLPPSGGDGRQGYRRP VSHPQRRRPTSEHALQEAKSNAPSSLRFKSENIVSYELEMPCRTTGEYRTPRVSASLQAVLAQAMQTGGDDIDIVDSHTNSLRSRLENIINAN SSSNGGPGSGAGPLAANTAGSGVGSMPNVRNIKSSRGLPSAGTALDSNRRPPTGRIPAKKPASAYPKARGLVNK</br/>//seq> Analyze the provided protein amino acid sequence, accurately predict and describe its function, subcellular_location, family, key_sequence_motif'\nResponse in the following exact format'\nFUNCTION:
...\nSUBCELLULAR LOCATION: ...\nFAMILY: ...\nKEY SEQUENCE MOTIF: ...\nDetailed Analysis [\sigma owords): [in one brief paragraph of \leq 300 words, describe potential catalytic mechanisms, key domains, post-translational modifications, associations with diseases, etc.]\nKeep the tone professional yet fluent; let every word earn its place. FUNCTION: Plus-end directed microtubule motor involved in anterograde axonal transport, potentially moving distinct cargos in fly neurons compared to other kinesin motors. inSUBCELLULAR LOCATION: Cytoplasm, cytoskeleton (microtubule-associatedy)nFAMILY: TRAFAC class myosin-kinesin ATPase superfamily, Kinesin family, Kinesin II subfamilynKEY SEQUENCE MOTIF: ATP-binding site (IOE-13)ynin****Extended Information***unkTipRoS1 is a microtubule-associated motor protein that utilizer Phydrolysis to drive anterograde transport along microtubules, likely facilitatin cargo movement in *Drosophila* neurons. It contains a conserved kinesin motor domain (19-344) responsible for ATP binding and microtubule interaction, along with two coiled-coil regions (350-384, 432-880) that may mediate dimerization or cargo binding. The protein exhibits multiple disordered regions and composinal biases, including basic and acidic residue clusters, which may regulate its function or interactions. RIp68D belongs to the Kinesin II subfamily, suggesting specialized roles distinct from conventional kinesin heavy chains. Its localization to the cytoskeleton underscores its involvement in microtubule-based modify moreogened. based motility processes. G-protein coupled receptor moody: <seq-MSDETTGSLGDAFSPMDTPTTTIMPPPADVDESGFSHSLLTFAAVMTFLIMIVGICGNLLTVVALLKCPKVRNVAAAFIISLCIADLLFCALVLPFQGLRFVQGTWRHGEVLCRLIPFIQYGNIGVSLLCIAMITINRYVMITHYSLYNRIYKRHWIAIMIAACWLFSY AAAFIISLCIADLIFCALVLPFGGLRFVQGTWRHGEVLCRLIPFIQYGNIGYSLCIAMITINRYVMITHYSLYNRIYKRHWIAIMIAACWLFSY
GMQLPTLIGAWGRFGYDARLQPTCSIMSDRHGHSSKTTIFITAFVIPCLVIIACYKIFWVYHKSEQRLKRHATKQNSIPNNLRPLAAATSMPS
GDGANPSQVPAGCRVSSDSSNYSTDVPDTTPGGAGGGAGVKQQPSRVKDQREVRAKRNEWRITKMVLAIFLSFVICYLPITIVKVADKDVE
HPSLHIFSYIMLYLSACINPIIVYIMKQYRKAYKTVVFCQPAARLLMPFGKGROSSAAEKWKDTGLSNNHSRTIVSQMSAGATATATATA
AGTQPQSTSTQGPVQALELTARVPDLISKSSNLPIPQPIPQIPPAGARPSLTPPPPPPSVLTATHSNGSGSQRLPLKKNNHSYTNSGFNSSVISAN
PSSSPSPSSGGGIVRPGIGSMGNGSASIRRITMVGDDIILEEEELPPTPTASSPPQMQAPPPPPSSSRQTTMNALNTTPKTPIYMNVDSPKRNQ
SYSERNIPVPAREGHDQGVKDSQGLPSKLMDKKKFPKD</bd> Analyze the provided protein amino acid sequence, accurately predict and describe its function, subcellular_location, family, key_sequence_motif\nResponse in the following exact format\nFUNCTION: ...\nSUBCELLULAR LOCATION: ...\nFAMILY: ...\nKEY SEQUENCE MOTIF: ...\nDetailed Analysis (≤ 300 words): [in one brief paragraph of ≤ 300 words, describe potential catalytic mechanisms, key domains, post-translational modifications, associations with diseases, etc.]\nKeep the tone professional yet fluent; let every word earn its place." Answer FUNCTION: Plus-end directed microtubule motor involved in anterograde axonal transport, potentially moving distinct cargos in fly neurons compared to other kinesin motors.'nSUBCELLULAR FUNCTION: Cytoplasm, cytoskeleton (microtubule-associated)'nFAMILY: TRAFAC class myosin-kinesin ATPase superfamily, Kinesin family, Kinesin II subfamily'nKEY SEQUENCE MOTTE-ATP-binding site (106-113)'n'n*Extended Information**'n'nKlp68D is a microtubule-associated motor protein that utilizes ATP hydrolysis to drive anterograde transport along microtubules, likely facilitating cargo movement in *Drosophila* neurons. It contains a conserved kinesin motor domain (19-344) responsible for ATP binding and microtubule interaction, along with two colled-coil regions (350-384, 432-589) that may mediate dimerization or cargo binding. The protein exhibits multiple disordered regions and compositional biases, including basic and acidic residue clusters, which may regulate its function or interactions. Klp68D belongs to the Kinesin II subfamily, suggesting specialized roles distinct from conventional kinesin heavy chains. Its localization to the cytoskeleton underscores its involvement in microtubule-based motility processes. Examples of true or false protein-QA seq>MAKRSSLYIRIVEGKNLPAKDITGSSDPYCIVKVDNEPIIRTATVWKTLCPFWGEEYOVHLPPTFHAVAFYVMDEDALSRDDVIGKVCLT Determine whether this statement about the given protein is true or false: although this calcium-binding protein translocates to the plasma membrane upon intracellular calcium elevation to inactivate Ras signaling, its pleckstrin homology domain mediates this membrane association through specific phosphoinositide interactions, which is essential for its GTPase-activating function. Answer False. The PH domain lacks phosphoinositide binding activity due to a leucine substitution at position 592, preventing it from mediating membrane association; calcium-dependent membrane binding occurs through its C2 domains instead. Protein -seep-MANQVNGNAVQLKEEEEPMDTSSVTHTEHYKTLIEAGLPQKVAERLDEIFQTGLVAYVDLDERAIDALREFNEEGALSVLQQFKESDLS HVQNKSAFLCGVMKTYRQREKQGSKVQESTKGFDEAKIKALLERTGYTLDVTTGQRKYGGPPPDSVYSGVQFGIGTEVFVGKIPRDLYEDE LVPLFEKAGPIWDLRLMMDPLSGQNRGYAFITFCGKEAAQEAVKLCDSYEIRPGKHLGVCISVANNRLFVGSIPKNKTKENILEEFSKVTEGL VDVILYHQPDDKKKNRGFCFLEYEDHKSAAQARRRLMSGKVKVWGNVVTVEWADPVEEPPPEVMAKVKVLFVRNLATTVTEEILEKSFSEF GKLERVKKLKDYAFVHFEDRGAAVKAMDEMNGKEIEGEEIEIVLAKPPDKKRKERQAARQASRSTAYEDYYYHPPPRMPPPIRGRGRGGGR GGYGYPPDYYGYEDYYDDYYGYDYHDYRGGYEDPYYGYYDDGYAVKGRGGRGGRGGRGAPPPPRBGRAGYSQRGAPLGPPRGSRG RGGPAQQQRGGSGSGSGNGGGNGGKRKADGYNQPDSKRQTNNQQNWGSQPIAQQPLQQGDYSGNYGYNNDNQEFYQDTYGQQW Determine whether this statement about the given protein is true or false: although this protein's SUMOylation at specific lysine residues supports its role in mRNA processing complexes, and mutations causing an autosomal recessive neurodevelopmental disorder implicate its functional importance, structural studies reveal that these covalent modifications occur exclusively at lysine-359 within its third RNA recognition motif domain False; Explanation: SUMOylation occurs at Lys-13, Lys-171, and Lys-359, not exclusively at Lys-359, and the associated neurodevelopmental disorder is autosomal dominant, not recessive.

Figure 8: Illustrative examples of the four bilingual QA types (Part 1).

Examples of descriptive text protein-QA Protein <eq>-Maspfffvfllsalslentyaspnyrealsksllffqqqrsgrlpsdqqlswrsssglsdgssahvdltggyydagdnvkfnfpmafttt mlswssleygkkmgpelqnsrvairwatdyllkcaratpgklyvgvgppngdhkcwerpebmdtprtysysyspsnpgsdvaaetaaalaas smyrrkvdpkysrlllatakkkwmqfaiqyrgaysnslsssvcpfycsysgykdellwgaawlhratndpyytnfikslgggdqpplfswdn KYAGAYVILSRRAVINK DNNFELYK OAAENEM CKILPNSPSSSTKYTK GGLMYK LPOSNLOYVTSITFILTTYAKYMKSTKOTEN CGI NALINLSKROVDYVLGVNPMKMSYMVGFSSNFPKRIHHRGSSLPSRAVRSNSLGCNGGFOSFRTONPNPNILTGAIVGGPNONDEYPDORD DYTR SEPATYINAA EVGPLAYEAA SR SP </br> Question In one coherent paragraph, describe the protein represented by the given amino-acid sequence, covering: (1) its main function and any bound cofactors/ligands, (2) subcellular localization and the biological proc olved, (3) key domains or motifs, and (4) the specific reaction catalyzed if it is an enzyme A brief overview of the protein with the provided amino acid sequence (Q2V4L8, GUN3_ARATH) is as follows: This enzyme, an endoglucanase 3 (EC 3.2.1.4) from *Arabidopsis thaliana*, catalyzes the A brief overview of the protein with the provided amino acid sequence (Q2V4L8, GUNS_ARA1H) is as tollows: This enzyme, an endoglucianase 3 (EC. 3.2.1.4) from "Arabidopsis thaliana", catalyzes the endohydrolysis of (1—4)-B-Qlucosidic linkages in cellulose, (lichenin, and cereal B-D-glucans, facilitating polysacide degradation without requiring bound cofactors. It is secreted to the extracellular region, particularly localized to the plant-type cell wall and Golgi apparatus, where it participates in cell wall organization and cellulose catabolic processes. Key features include its classification within the glycosyl hydrolase 9 family, a catalytic nucleophile at position 77, additional active sites at residues 402, 453, and 462, an N-linked glycosylation site at Asn-370, and an N-terminal signal peptide (residues 1–21) for secretion. Functionally, it mediates root cap cell sloughing through cell wall breakdown, is specifically expressed in root cap cells, and is down-regulated by auxin and abscisic acid. The mature chain spans residues 22–484 and may undergo alternative splicing. Protein seq>MISTLRRAVFLRRFPAVVSPIKRAFSSRIDDEFDPQIMNINELNQEMQSIFGQEPSPDGPGTMDFSELKSSKIEPLRSKNIDFRQQIEYHKS THSSKNDSQAIEGYAKVASDMSKLTHVGIAGEAQMYDVSSKDNSKRTALACCKVILGKRVFDLVLANQMGKGDVLGVAKIAGINGAKQTSS LIPLCHNIALTHVRVDLRLNPEDFSVDIEGEASCTGKTGVEMEAMTAVSVAGLTVYDMCKAASKDISITDVRLERKTGGKSGSWSRL</br> Determine whether this statement about the given protein is true or false; although this protein's SUMOylation at specific lysine residues supports its role in mRNA processing complexes, and mutations causing an autosomal recessive neurodevelopmental disorder implicate its functional importance, structural studies reveal that these covalent modifications occur exclusively at lysine-359 within its third RNA recognition motif domain A concise description of the protein with the specified amino acid sequence (Q39056, CNX3_ARATH) is as follows:\n\n\text{This mitochondrial protein, localized specifically within the mitochondrial matrix via an N-terminal transit peptide (residues 1-32), functions as a cyclic pyranopterin monophosphate synthase (EC 4.6.1.17), It is essential for molybdenum cofactor biosynthesis, catalyzing the specific lyase reaction that converts (8S)-3',8-cyclo-7,8-dihydroguanosine 5'-triphosphate into cyclic pyranopterin monophosphate (cPMP) and diphosphate (RHEA:49580). Key residues form an active site (including residue 240) and substrate binding sites (residues 187-189 and 225-226). Belonging to the MoaC protein family, it assembles as a homohexamer and is predominantly expressed in the roots of *Arabidopsis thaliana*. Its primary biological role is in the \"Mo-molybdopterin cofactor biosynthetic process\" (GO:0006777). Examples of knowledge-based protein-QA Seep-MSAAILGQVWTRKLLPIPWRLCVPGRCVSSNFKAADLQVQVTREPQKKPAPSQPLLFGKTFTDHMLMVEWNSKTGWGPPRIQPFQNLTL HPACSGLHYSLQLFEGLKAYKGRDKQVRLFRPWLNMDRMLRSARRLCLPDFDKQELLECIRQLIEVDKDWVPDGNGTSLYVRPVLIGNEPS LGVGMYTQALLFVILCPVGSYFPGDSMTPVSLLADPSFVRAWIGGVGDCKLGGNYGPTVAVQQEAQKKGCEQVLWLYGPDHQLTEVGTMMI FVYWTHEDGELELATPPLDGIILPGVVRQSLLDLARTWGEFRVAERKVTMKELKRALEEGRVREVFGSGTACQVCPVHQILYEGKQLHIPTM ENGPELILRFOKELKAIOYGTSAHDWMLRV< You are a molecular biology professor writing a protein description for undergraduates. Based on the characteristics of the given protein amino acid sequence, briefly answer the following questions: 1. What is the primary catalytic function of this protein in branched-chain amino acid metabolism/n2. Where is this protein localized within the cell?'n3. What cofactor is essential for its enzymatic activity' via H. How does its substrate affinity (KM) vary among the branched-chain amino acids/m5. What is the quaternary struct of this protein?m0. Which post-translational modifications are known to occur on this protein?'n7. What is the physiological direction of its catalytic reactions?'n8. In which tissues is this protein expressed?'n9. What family of enzymes does this protein belong to?

1. It catalyzes the transamination of leucine, isoleucine, and valine with 2-oxoglutarate to form their respective α-keto acids and glutamate \n2. It is localized in the mitochondrion \n3. Pyridoxal 5'-phosphate is the essential cofactor\n4. The KM values are 1 mM for L-leucine and L-isoleucine, and 5 mM for L-valine\n5. It forms a homodimer\n6. It undergoes N6-(pyridoxal phosphate)lysine modification at position 230 and N6-acetyllysine at position 320 and

<seq>MAFSMLFTGHAECVKFVKKFNLPLLVTGGGGYTKENVARCWTVETGILLDTELPNEISENDYIKYFAPDFSLKIPGGHIENLNTKSYISSI KVQILENLRYIQHAPSVQMQEVPPDFYIPDFDEDEQNPDVRVDQRSRDKQIQRDDEYFDGDNDNDAS</seq>

Question

You are a molecular biology professor writing a protein description for undergraduates. Based on the characteristics of the given protein amino acid sequence, briefly answer the following questions: 1. What is the enzymatic reaction catalyzed by this protein? Which is protein localized within the cell?na3. What is the role of this protein in chromatin organization?n4. Which histone proteins substrates for its deacetylation activity?n5. How does it sactivity contribute to transcriptional regulation?n6. What are functional implications of its disordered region (residues 120-185)?n7. Does it form part of any multiprotein complexes?n8. What is the significance of its basic and acidic residue compositional biases?n9. How does it compare to other members of the histone deacetylase family?

Answer

1. It catalyzes the hydrolysis of N(6)-acetyl-L-lysyl-[histone] to produce L-lysyl-[histone] and acetate.\n2. It is localized in the nucleus.\n3. It is involved in chromatin organization through histone deacetylation, leading to epigenetic repression.\n4. It deacetylates lysine residues on the core histones H2A, H2B, H3, and H4\n5. Its deacetylation activity represses transcription by modifying chromatin structure.\n6. The disordered region may contribute to protein flexibility or interaction with other molecules.\n7. Yes, it acts via the formation of large multiprotein complexes.\n8. These biases may influence protein interactions or substrate binding.\n9. It belongs to the histone deacetylase family and shares similar functional domains.

Figure 9: Illustrative examples of the four bilingual QA types (Part 2).

Prompt for Knowledge-based QA generation is following:

"Based on the provided annotations, generate exactly 1-9 distinct, single-sentence questions that a researcher would naturally ask to fully interrogate this protein. Guidelines:

(1) Each question must probe a different biological dimension (expression, localization, mechanism, regulation, phenotype, disease, evolution, interaction, structure/properties).

(2) Keep questions concise, fluent.

- (3) One per line, numbering, and the corresponding answers to these questions are concise and clear.
- (4) The questions can be appropriately flexible and occasionally combined with some actual scenarios or content related to species.
- The Questions and Answers are divided into two parts (wrapped with <Questions><\\Questions> and <Answers><\\Answers> respectively). All the information in the Q&A should be based entirely on the given annotations and should not be supplemented by yourself."