Protein2Text: Resampling Mechanism to Translate Protein Sequences into Human-Interpretable Text

Ala Jararweh*,1,2, **Oladimeji Macaulay***,2, **David Arredondo**2, **Yue Hu**2, **Luis Tafoya**2,
Kushal Virupakshappa2, **Avinash Sahu**1,2
[1]Department of Computer Science, The University of New Mexico
[2]Comprehensive Cancer Center, The University of New Mexico
`{ajararweh,asahu}@salud.unm.edu`

Abstract

Proteins play critical roles in biological systems, yet 99.7% of over 227 million known protein sequences remain uncharacterized due to the limitations of experimental methods. To assist experimentalists in narrowing down hypotheses and accelerating protein characterization, we present Protein2Text, a multimodal large language model that interprets protein sequences and generates informative text to address open-ended questions about protein functions and attributes. By integrating a resampling mechanism within an adapted LLaVA framework, our model effectively maps protein sequences into a language-compatible space, enhancing its capability to handle diverse and complex queries. Trained on a newly curated dataset derived from PubMed articles and rigorously evaluated using four comprehensive benchmarks—including in-domain and cross-domain evaluations—Protein2Text outperforms several existing models in open-ended question-answering tasks. Our work also highlights the limitations of current evaluation metrics applied to template-based approaches, which may lead to misleading results, emphasizing the need for unbiased assessment methods. Our model weights, evaluation datasets, and evaluation scripts are publicly available at `https://github.com/alaaj27/Protein2Text.git`.

1 Introduction

Proteins are essential to nearly all biological processes. Understanding protein functions is crucial for unraveling disease mechanisms, predicting the effects of genetic mutations in conditions like cancer, and discovering targeted and personalized therapeutics (Liu et al., 2020; Quazi, 2022; Wu et al., 2023b). Despite the characterization of 460,000 proteins in UniProt (Consortium, 2022), a staggering 99.7% of the 227 million protein sequences remain poorly characterized (Consortium, 2022; Coudert et al., 2022). This vast number of uncharacterized proteins poses a significant bottleneck in biomedical research, impeding the full realization of the potential envisioned with the sequencing of the human genome. Experimental methods for protein characterization are inherently time-consuming and costly, making it impractical to scale to millions of proteins. Therefore, there is an urgent need for computational methods to complement and accelerate traditional experimental approaches.

For the first time, Large Language Models (LLMs) are offering an alternative to these challenges. For example, encoder-based models like ESM-2 and OntoProtein leverage masked language modeling on protein sequences to generate embeddings that capture structural and functional information (Lin et al., 2022b,a; Zhang et al., 2023, 2022). Similarly, to predict gene/protein structural and functional information, several approaches use other modalities such as text (Jararweh et al., 2024) and expression (Du et al., 2019; Cui et al., 2024). Decoder-based models such as AlphaFold predict 3D structures from amino acid sequences (John Jumper and Hassabis, 2021). Moreover, multimodal LLMs have been developed to bridge the gap between biological sequences and natural language, translating complex protein data into accessible human language (Luo et al., 2023; Fang et al., 2024). Bimodal Protein Language Models (PLMs), including ProteinChat and ProtChatGPT (Guo et al., 2023; Wang et al., 2024), attempt to co-embed protein sequences with natural language using projection mechanisms.

However, existing PLMs face limitations. A critical gap is the lack of rigorous quantitative evaluation on question-answering (QA) tasks, which are vital for practical utility. Many PLMs depend on template-based QA datasets, transforming structured data into unstructured text using

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 918–937
April 30, 2025 ©2025 Association for Computational Linguistics

fixed templates (Guo et al., 2023; Xiao et al., 2024a; Luo et al., 2023). This methodology limits the models' ability to generalize to new, unseen queries and diminishes their adaptability to diverse instructions. Consequently, template-based QA datasets hinder model expressiveness, and often –as we also demonstrate – overfit to specific patterns and lack the conversational flexibility necessary for addressing complex research questions (see Table 14) (Liu et al., 2024).

Therefore, we present a novel multimodal reasoning model that modifies the LLaVA (Liu et al., 2023a) framework to adopt for the protein domain. Our model provides real-time, interactive analysis of protein properties and handles complex, open-ended questions, empowering researchers to gain actionable insights for laboratory research. Trained on a newly curated dataset derived from published literature on proteins in PubMed articles, our model benefits from a rich and diverse corpus surpassing template-based methods' limitations. We also compiled four comprehensive evaluation datasets to benchmark our model against existing PLMs rigorously. By releasing these evaluation datasets and model weights, we aim to promote a thorough assessment of protein LLMs across a wide range of tasks and specialized datasets.

2 Related Work

The sequential nature of protein primary structure lends itself to language modeling for protein characterization. For example, encoder-based LLMs trained on protein amino acid sequences have been adopted to generate a representation space that captures sequence structures (Lin et al., 2022b,a; Elnaggar et al., 2021; Zhang et al., 2022). Generative LLMs have also been proposed for a variety of protein generation tasks such as generating 3D structure (John Jumper and Hassabis, 2021), and novel protein sequences (Madani et al., 2020; Nijkamp et al., 2022; Lv et al., 2024). LLMs that incorporate natural language and protein as one modality (i.e. considering protein as text modality) have been proposed. For example, Galactica models are general-purpose LLMs that are trained on scientific corpora to perform different reasoning tasks including protein captioning. Several studies attempt to integrate text with protein modalities such as DNA/RNA sequences (Richard et al., 2024), 3D structure (Guo et al., 2023; Wang et al., 2024), and amino acid sequences (Xiao et al., 2024b; Luo et al., 2023). Similarly, multi-modality projection similar to vision-language alignment (Alayrac et al., 2022; Liu et al., 2023a), has been applied to align between protein and natural text where protein is considered as single modality (Guo et al., 2023; Wang et al., 2024; Liu et al., 2024; Luo et al., 2023; Fang et al., 2024). See Appendix D for detailed discussion on related work.

3 Protein2Text

Protein Encoder. Our approach is based on LLaVA (Liu et al., 2023a) which integrates images and text via instruction tuning. We adopt LLaVA to protein amino acid sequences by replacing the image encoder with a protein encoder (Figure 1b). We use ESM-2 (Lin et al., 2022b) a transformer-based encoder that has 33 transformer layers and a total of 652 million parameters. Every sequence ($\mathcal{P}$) is encoded to a multidimensional token embedding using ESM2 (ϕ_{esm}) where every character is considered a token. Formally:

$$\mathbf{Z}_v = \phi_{esm}(\mathcal{P})$$

where $\mathbf{Z}_v \in \mathbb{R}^{d\times T_1}$ represents the embedding of the protein tokens where d is the dimension size and T_1 is the number of tokens.

LLM Encoder. Simultaneously, the instruction/question $\mathbf{X}_q$, given as natural language input, is tokenized and embedded using LLaMA-3, ϕ_{LLM}:

$$\mathbf{H}_q = \phi_{LLM}(\mathbf{X}_q)$$

where $\mathbf{H}_q \in \mathbb{R}^{k\times T_2}$ represents the token embeddings of the instruction, with k being the embedding dimension and T_2 the number of tokens.

Perceiver Resampler. In LLaVA, images are divided into a fixed number of patches, yielding a fixed number of image tokens without losing information. However, protein sequences have different sizes, and truncating them to a fixed size might remove potentially critical information. To this end, we extend their architecture by adding a protein resampler (Jaegle et al., 2021; Carion et al., 2020; Alayrac et al., 2022). The resampler finds a fixed number of latent tokens from varying-size protein sequences (Figure 1b). This reduces the computational complexity of the cross-attention in the LLM and prevents long protein tokens from

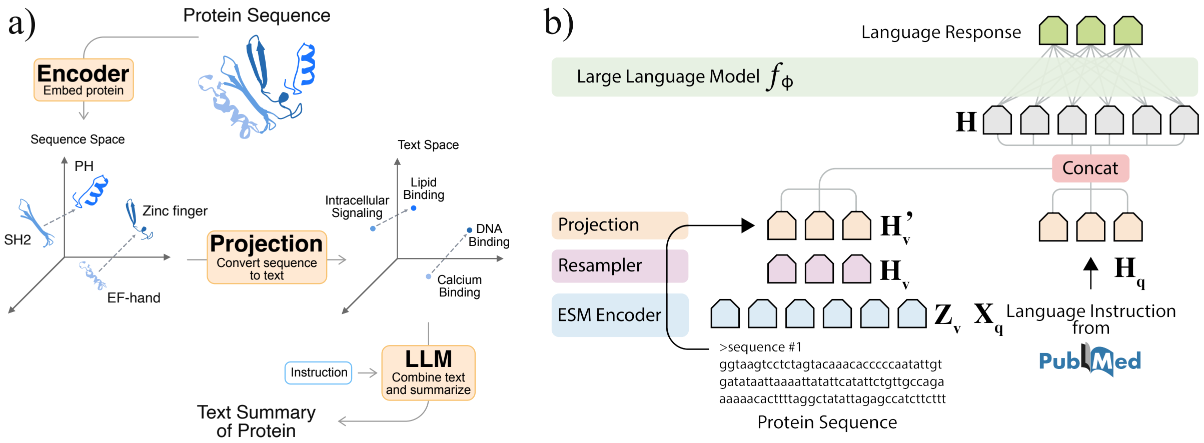


Figure 1: **Protein2Text Architecture Overview.** a) Protein2Text generates descriptive text from amino acid sequences by combining pre-trained protein encoder and language models. b) The protein tokens are compressed into latent tokens using the resampler and projected to the language space using the projector.

exhausting the model's maximum length. Given the protein token embeddings ($\mathbf{Z}_v$), the resampler generates $\mathbf{H}_v \in \mathbb{R}^{d \times S}$, where S is the number of latent tokens that compress the information in the original tokens:

$$\mathbf{H}_v = \phi_{Resampler}(\mathbf{Z}_v)$$

Protein2Text Projector. To align the protein and the text modalities (Figure 1a), we project the dimensions of protein latent tokens (d) into the language embedding space (k) via the projector:

$$\mathbf{H}'_v = W \cdot \mathbf{H}_v$$

where W is the set of trainable parameters and $\mathbf{H}'_v \in \mathbb{R}^{k \times S}$. The projected tokens are then concatenated to the text tokens, producing $\mathbf{H} \in \mathbb{R}^{k \times (S+T_2)}$. $\mathbf{H}$ is then fed to the LLM decoder (f_ϕ) to generate the response.

Dataset Collection We collect four different datasets tailored to distinct requirements. First, the pretraining dataset spans 394,000 protein amino acid sequences and function descriptions collected from UniProt (Consortium, 2022). This dataset is entirely used to train the resampler and the projector during the pretraining stage.

Next, we generate a comprehensive question and answering dataset (i.e. **Protein2Text-QA**) to fine-tune the model parameters. The dataset spans approximately 210,000 pairs of QA. We utilize research carried out on proteins from published articles in the PubMed Central (PMC) database (Consortium, 2015) to create questions and answers. Articles mentioning the protein names are located and fed to the LLaMA3.1 model to generate a series of QA pairs, such that they focus only on the protein name given.

The test set and zero-shot set are then sampled from the Protein2Text-QA dataset. The proteins in the test set can be found in the pretraining dataset but not in the fine-tuning dataset. On the other hand, the zero-shot set is sampled such that the protein sequences and their variants are not mentioned in both pre-training and fine-tuning sets. The variants were also filtered out to eliminate data leakage (Bushuiev et al., 2024) since some protein variants might have different sequences but similar/same function (Brett et al., 2002; Schlüter et al., 2009). Finally, we generate two cross-domain datasets to evaluate the model on questions not mentioned in the abstracts. First, the **DiscussionQA** which spans QAs extracted from discussion sections, and the **IntroductionQA** which spans QAs extracted from introduction sections. The collection process of training and evaluation datasets, and detailed statistics, generation pipelines, preprocessing, and sample QAs are further discussed in Appendix A.

Training. The model training process consists of two stages: pretraining and fine-tuning. During pretraining, we freeze the protein encoder and the LLM while the parameters for the resampler and projector are trained. Next, we perform fine-tuning, where we train the entire model except the protein encoder parameters. In this stage, the LLM is trained using Low-Rank Adaptation (LoRA) (Hu et al., 2021). Finally, we assess the performance by designing four evaluation datasets tailored to distinct requirements such as baseline benchmarking, zero-shot ability, and cross-domain evaluations. Further details about training details, hyperparameters, baselines, and benchmarks are discussed in Appendices C.1, C.2, E, and F respectively.

4 Experiments

4.1 Protein2Text-QA Evaluation

Experiment. We evaluated the performance of Protein2Text against two categories of large language models (LLMs): general-purpose LLMs and protein-specific LLMs. For general-purpose LLMs, such as GPT4o-mini (OpenAI et al., 2023) and LLaMA3.1 (Dubey et al., 2024), the evaluation focused on assessing the degree of potential data leakage within the question prompts. We hypothesized that if the answers were embedded in the question prompts, general-purpose LLMs would likely respond correctly (Cadene et al., 2020). In the second category, we benchmarked Protein2Text against multimodal LLMs tailored for protein-related tasks, including Mol-Instruction (Fang et al., 2024), BioMedGPT (Luo et al., 2023), and ProtT3 (Liu et al., 2024), all of which are open-source tools. We evaluated the performance using lexical metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), and semantic similarity metrics such as BERT similarity (Devlin et al., 2019), and BiomedBERT similarity (Gu et al., 2021). Further details on baseline models and scores can be found in Appendices E and G.

Findings. Table 1 summarizes the performance of models on the Protein2Text QA test set. General-purpose LLMs exhibited poor performance due to their inability to interpret protein sequences (see Table 6), indicating minimal data leakage from the prompts. In contrast, protein-specific LLMs like BioMedGPT and Mol-Instruction showed competitive performance likely because they are also trained on PubMed data. BioMedGPT achieved higher semantic similarity scores but lower lexical scores compared to Mol-Instruction, suggesting its answers were semantically relevant but not necessarily accurate (Table 6). ProtT3, trained on template-based benchmarks or short QA (1–3 words), struggled with out-of-domain instructions, unlike Protein2Text, Mol-Instruction, and BioMedGPT.

Protein2Text consistently outperformed baselines across both semantic and lexical metrics. To explore potential enhancements, we implemented a Gated cross-attention (GCA) mechanism (Jia et al., 2024; Das et al., 2022; Alayrac et al., 2022) at the top of the resampler architecture. Surprisingly, adding GCA resulted in reduced performance; therefore, was excluded in the final Protien2Text. Further investigation is needed to determine whether this decrease is due to the increased number of parameters requiring larger training data or if GCA is ill-suited for this problem. Details on parameter counts and the GCA ablation study are provided in Table 9 and Table 8, respectively.

4.2 Cross-domain Evaluations

Experiment. We assess Protein2Text's generalizability to new domains. Here, we evaluate the performance on the zero-shot QA where proteins and their variants in this set are hidden during the entire training pipeline. Similarly, we assess the performance where the domain of the extracted QA is different such as the introduction (**IntroductionQA**) and discussion (**IntroductionQA**) sections. We focus on the PLM baselines throughout this experiment due to their superior performance compared to general-purpose LLMs.

Findings. First, the baselines showed similar performance in the Zero-shotQA (Table 2) compared to their performance in the test set (Table 1). Even though our model matches and often outperforms the baselines, the performance expectedly dropped compared to the test set. Since proteins and their variants were hidden during the alignment stage, novel sequence domains might have been introduced, hindering the resampler compression. The baselines showing similar performance could also indicate that these proteins might have been seen by these models during their training. Next, we evaluate the performance on the **IntroductionQA** as demonstrated in Table 3. Our model outperforms the baselines across lexical and semantic metrics; however, we see a slight decrease in metrics performance compared to the test set. This is likely because introduction sections usually present new information that was not necessarily mentioned in the abstract (Cohen et al., 2010). For the **DiscussionQA**, however, we found the performance of QAs from abstracts is similar

Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	BiomedBERT Score
General-purpose LLMs								
GPT4o-mini	0.0202	0.0088	0.0698	0.0279	0.0589	0.156	0.67	0.88
LLaMA3.1	0.0137	0.0067	0.0422	0.0186	0.0387	0.1100	0.613	0.8014
Protein-specific LLMs								
BioMedGPT	0.074	0.035	0.160	0.056	0.144	0.140	0.750	0.905
Mol-Instructions	0.065	0.036	0.187	0.092	0.168	0.273	0.743	0.878
ProtT3	6×10^{-6}	1×10^{-6}	0.062	0.001	0.061	0.0174	0.768	0.843
Protein2Text	**0.144**	**0.083**	**0.322**	**0.18**	**0.288**	**0.377**	**0.891**	**0.943**

Table 1: **Baseline comparison on our Protein2TextQA test set**. **Bold** and underline denote best and second best performing models respectively. For all metrics, higher values indicate better performance.

Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	BiomedBERT Score
BioMedGPT *	**0.075**	0.0347	0.159	0.0536	0.1429	0.139	0.750	0.905
Mol-Instructions*	0.067	**0.038**	0.193	0.0953	0.172	0.282	0.744	0.880
ProtT3 *	7×10^{-6}	9×10^{-7}	0.062	0.001	0.061	0.017	0.769	0.843
Protein2Text	0.043	0.0248	**0.265**	**0.148**	**0.239**	**0.326**	**0.815**	**0.897**

Table 2: **Zero-shot analysis on unseen proteins.** Proteins and their variants, in this analysis, were held out during the two stages of Protein2Text training. However, it is not guaranteed that these proteins were also hidden during the training of the baselines (i.e. denoted by *).

to the performance of those extracted from the discussion sections as shown in Table 4, suggesting that discussion and abstract sections are more semantically aligned.

4.3 ProteinKG25 Benchmark Evaluation

The ProteinKG25 dataset, originally designed as a knowledge base for protein attributes, was adapted into a question-answering (QA) format using templated questions by the authors of ProtT3 (Liu et al., 2024) (see Appendix F). ProtT3 was fine-tuned specifically on this templated dataset. We evaluated our Protein2Text model on this benchmark in a zero-shot manner, without any additional fine-tuning.

As anticipated, ProtT3 achieved high-performance metrics on lexical evaluation scores (Table 13). However, we observed that in template-based scenarios, these metrics might not fully capture a model's ability to predict embedded protein attributes in the template. Models trained on templates can replicate the template structure, leading to high lexical similarity scores, even if the critical details within the responses are incorrect. Using the empty template as the prediction and ignoring attributes in the blanks achieved high lexical scores (Table 13). In contrast, models like Protein2Text, which are not trained on these templates, may generate responses that deviate from the template format, resulting in lower performance despite potentially providing accurate and informative answers.

To investigate this further, we focused on the task of predicting protein subcellular localization, a classification problem present in the ProteinKG25 dataset. We specifically prompted the models to predict protein localization among three classes and assessed their outputs using standard classification accuracy.

Our results indicated that while the template-trained models achieved high lexical similarity metrics (Table 13), they exhibited lower classification accuracy on the protein localization task (Figure 2a). This suggests that these models, despite effectively reproducing the template structure, may not reliably predict the correct protein attributes. In contrast, Protein2Text demonstrated higher classification accuracy in this task, indicating a better ability to generalize and accurately predict protein localization in a zero-shot setting. Furthermore, we observed that the LitGene-based encoder predictor, which was specifically fine-tuned for protein localization, achieved the highest accuracy among the models evaluated. It suggests that decoder-based models like Protein2Text would benefit from further enhancements, such as larger or more diverse training datasets or architectural improvements, to close the performance gap, as GPT-4 and other general-purpose LLMs have matched supervised models for general NLP tasks.

Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	BiomedBERT Score
BioMedGPT	0.068	0.032	0.172	0.059	0.152	0.133	0.754	0.907
ProtT3	5×10^{-6}	6×10^{-159}	0.054	0.001	0.052	0.0167	0.748	0.840
Mol-Instructions	0.072	0.042	0.196	0.099	0.17079	0.287	0.733	0.877
Protein2Text	**0.130**	**0.078**	**0.318**	**0.181**	**0.279**	**0.366**	**0.882**	**0.939**

Table 3: **Model evaluation on the IntroductioQA set.** The QA dataset is constructed from article introductions.

Model	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	BiomedBERT Score
BioMedGPT	0.0577	0.0272	0.1724	0.0601	0.1506	0.1316	0.7344	0.9057
Mol-Instructions	0.0795	0.0475	0.2135	0.1159	0.1892	0.0475	0.743	0.8771
ProtT3	2×10^{-6}	3×10^{-7}	0.05407	0.00166	0.05276	0.015878	0.7465	0.8387
Protein2Text	**0.143**	**0.089**	**0.346**	**0.212**	**0.311**	**0.392**	**0.895**	**0.943**

Table 4: **Evaluating Protein2Text on the DiscussionQA set.** The DiscussionQA set is constructed from discussion sections of PubMed articles.

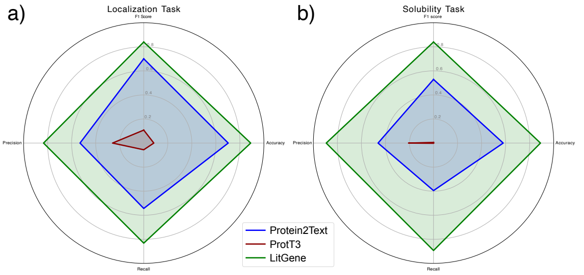


Figure 2: Evaluation on protein attribute prediction tasks: a) Subcellular localization and b) Protein solubility.

We extended our evaluation to protein solubility prediction tasks and observed similar trends. The template-trained models again showed high lexical similarity scores but lower classification accuracy compared to Protein2Text and the fine-tuned encoder-based model (Figure 2b). These findings reinforce the notion that while template-based models excel in reproducing specific formats, they may not always capture the underlying protein attributes accurately.

4.4 Ablation Study

Since wide-range ablation studies are prohibitive and time-consuming in LLMs due to their training time, we focus on more targeted ablation such as the extension of our model beyond LLaVA, the resampler. To assess the effect of the resampler, we remove it from the model. In this case, the <CLS> token from the protein encoder is used as the sole token representing the protein sequence, resulting in a single token projection. We compare this to the proposed model, in which the resampler creates 128 tokens, distilled from the embeddings of the entire protein sequence including the <CLS> token. The resampler uses roughly twofold the number of trainable parameters compared to the projector-only model (Table 9). We find a significant increase in performance across all metrics when moving from a single token projection to the latent 128 tokens generated by the resampler (Table 5).

4.5 Qualitative Analysis

The results in Table 6 and 15 show substantial differences in the ability of various baseline models to answer questions related to proteins accurately and contextually. Protein2Text outputs precise and detailed responses that align closely with the truth of the ground in most questions. For example, it effectively identifies the primary function of `P29474` and correctly notes the mitochondrial localization of `A0A1W2PPH1`. In contrast, models like ProtT3 and Mol-Instructions often produce incomplete or speculative answers, demonstrating limited generalization. Also, Protein2Text performed well in handling nuanced queries, such as the impact of nitrosylation on protein activity or changes in protein abundance at high altitudes. Nonetheless, despite its overall accuracy, Protein2Text occasionally fails and confidently generates incorrect answers (Table 15).

5 Conclusion

We introduced Protein2Text, a multimodal large language model designed to interpret protein sequences and generate descriptive text, facil-

Protein2Text	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	Biomed-BERT Score
-w/o resampler	0.054	0.029	0.249	0.134	0.224	0.301	0.818	0.917
-w/ resampler	**0.144**	**0.083**	**0.322**	**0.180**	**0.288**	**0.378**	**0.891**	**0.942**

Table 5: **Resampler ablation study.** The model highlighted in **bold** denotes the best-performing model.

Protein ID/ Question	**Ground Truth**	**Model Outputs**
ID: P29474 **Q1**: What is the primary function of this protein in the body?	This protein is involved in the synthesis of nitric oxide.	**Protein2text:** The primary function of this protein is related to the metabolism of nitric oxide. It is involved in the regulation of nitric oxide levels in the body, which is essential for various physiological processes. **LLaMA3:** Based on the provided sequence, it appears to be a protein sequence. However, without further information or context, it is challenging to determine the exact function of this protein in the body. **ProtT3:** IMMUNE SYSTEM **BioMedGPT:** May be involved in the regulation of cell proliferation. May be involved in the regulation of cell cycle progression. May be involved in the regulation. **Mol-Instructions:** The primary function of the protein MGNLKSV…LK
ID: A0A1W2PPH1 **Q2**: Where is this protein located in the cell?	It is located in the mitochondria.	**Protein2Text:** This protein is located in the mitochondria. **LLaMA3:** Upon analyzing the sequence, I notice that it contains a transmembrane domain, which is a characteristic of integral membrane proteins. **ProtT3:** This protein is involved in the following processs: signal transduction, detection of chemical stimulus involved in sensory perception of smell, G protein-coupled receptor signaling pathway, sensory perception of smell, and response to stimulus. This protein is located in the following components: cytoplasm, membrane, and nucleus. **BioMedGPT:** Mitochondrion **Mol-Instructions:** The protein MLSRLRVVSTALT…VM
ID: Q5T7M4 **Q3**: What happens to the levels of this protein in obese and insulin-resistant individuals?	The levels of this protein are reduced in obese and insulin-resistant individuals.	**Protein2Text:** The levels of this protein are lower in obese and insulin-resistant individuals. **LLaMA3:** In obese and insulin-resistant individuals, the levels are typically decreased. **ProtT3:** NO **BioMedGPT:** May be involved in the regulation of insulin sensitivity. **Mol-Instructions:** The levels of this protein are known to be altered in obese and insulin-resistant individuals. Specifically, studies have shown that obese individuals tend to have lower levels of this protein compared to non-obese individuals.

Table 6: **Qualitative analysis.** Comparing responses of baseline models on different protein-related questions. More examples can be found in Table 15.

itating complex and open-ended question-and-answer tasks related to protein functions and attributes. By employing a resampling mechanism to embed protein sequences into a human-interpretable space compatible with language models, Protein2Text demonstrated strong performance on various benchmarks, outperforming general-purpose and several protein-specific multimodal LLMs, particularly in open-ended QA tasks. The model showed robustness across different types of textual inputs derived from scientific literature in both fine-tuned and zero-shot settings. To enable rigorous benchmarking, we compiled four new datasets to evaluate in-domain and cross-domain capabilities. Our analyses also revealed limitations of current metrics when dealing with template-based datasets like ProteinKG25, indicating that standard lexical similarity metrics may not fully capture a model's ability to predict specific protein attributes and highlighting the need for cautious interpretation of these metrics. Incorporating task-specific fine-tuning or architectural adjustments may help bridge the gap between decoder-based models like Protein2Text and specialized encoder-based models in certain applications.

By providing a framework capable of interpreting protein sequences and generating informative text, our work demonstrates the potential to use multimodal language models for protein analysis, which may assist researchers in exploring protein functions and attributes. We hope that releasing our evaluation datasets and model weights will encourage further research and development in this area, ultimately contributing to advancements in computational biology and bioinformatics. Protien2text is not immune to occasional hallucinations of incorrect answers, which represents an important avenue for future work.

6 Acknowledgements

This project was partially supported by R00CA248953 and the UNM Comprehensive Cancer Center Support Grant NCI P30CA118100. This research also used resources of the Oak

Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE- AC05-00OR22725.

7 Ethical Considerations

AI has a major impact on the scientific, health, and social fields. We encourage responsible evaluation of LLMs to eliminate potential biases that could affect future applications. We also encourage the responsible usage of resource and utilizing Low-Rank fine-tuning mechanisms when applicable, aiming to alleviate environmental risk. Protein2Text is an AI agent that is meant to positively contribute to the current progress by advancing state-of-the-art results and providing new evaluation benchmarks. However, our evaluation indicates that the model occasionally outputs incorrect answers confidently when uncertainty is warranted. As such, Protein2Text should be used as a complementary tool, with its outputs critically assessed by experts who understand the model's limitations.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *Preprint*, arXiv:2310.10449.

Michael Benington, Leo Phan, Chris Pierre Paul, Evan Shoemaker, Priyanka Ranade, Torstein Collett, Grant Hodgson Perez, and Christopher Krieger. 2023. Scaling studies for efficient parameter search and parallelism for large language model pretraining. *Preprint*, arXiv:2310.05350.

David Brett, Heiko Pospisil, Juan Valcárcel, Jens Reich, and Peer Bork. 2002. Alternative splicing and genome complexity. *Nature Genetics*, 30(1):29–30. Epub 2001 Dec 17.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Anton Bushuiev, Roman Bushuiev, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. 2024. Revealing data leakage in protein interaction benchmarks. *Preprint*, arXiv:2404.10457.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. 2020. Rubi: Reducing unimodal biases in visual question answering. *Preprint*, arXiv:1906.10169.

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

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. *Preprint*, arXiv:2005.12872.

K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

Europe PMC Consortium. 2015. Europe pmc: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43(D1):D1042–D1048.

Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, Hill DP, Lee R, Mi H, Moxon S, Mungall CJ, Muruganugan A, Mushayahama T, Sternberg PW, Thomas PD, Van Auken K, Ramsey J, Siegele DA, Chisholm RL, Fey P, Aspromonte MC, Nugnes MV, Quaglia F, Tosatto S, Giglio M, Nadendla S, Antonazzo G, Attrill H, Dos Santos G, Marygold S, Strelets V, Tabone CJ, Thurmond J, Zhou P, Ahmed SH, Asanitthong P, Luna Buitrago D, Erdol MN, Gage MC, Ali Kadhum M, Li KYC, Long M, Michalak A, Pesala A, Pritazahra A, Saverimuttu SCC, Su R, Thurlow KE, Lovering RC, Logie C, Oliferenko S, Blake J, Christie K, Corbani L, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov D, Smith C, Cuzick A, Seager J, Cooper L, Elser J, Jaiswal P, Gupta P, Jaiswal P, Naithani S, Lera-Ramirez M, Rutherford K, Wood V, De Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Tutaj MA, Vedi M, Wang SJ, D'Eustachio P, Aimo L, Axelsen K, Bridge A, Hyka-Nouspikel N, Morgat A, Aleksander SA, Cherry JM, Engel SR, Karra K, Miyasato

SR, Nash RS, Skrzypek MS, Weng S, Wong ED, Bakker E, Berardini TZ, Reiser L, Auchincloss A, Axelsen K, Argoud-Puy G, Blatter MC, Boutet E, Breuza L, Bridge A, Casals-Casas C, Coudert E, Estreicher A, Livia Famiglietti M, Feuermann M, Gos A, Gruaz-Gumowski N, Hulo C, Hyka-Nouspikel N, Jungo F, Le Mercier P, Lieberherr D, Masson P, Morgat A, Pedruzzi I, Pourcel L, Poux S, Rivoire C, Sundaram S, Bateman A, Bowler-Barnett E, Bye-A-Jee H, Denny P, Ignatchenko A, Ishtiaq R, Lock A, Lussi Y, Magrane M, Martin MJ, Orchard S, Raposo P, Speretta E, Tyagi N, Warner K, Zaru R, Diehl AD, Lee R, Chan J, Diamantakis S, Raciti D, Zarowiecki M, Fisher M, James-Zorn C, Ponferrada V, Zorn A, Ramachandran S, Ruzicka L, and Westerfield M. 2023. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031.

The UniProt Consortium. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Elisabeth Coudert, Sebastien Gehant, Edouard de Castro, Monica Pozzato, Delphine Baratin, Teresa Neto, Christian J A Sigrist, Nicole Redaschi, Alan Bridge, and The UniProt Consortium. 2022. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics*, 39(1):btac793.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480.

Sowmen Das, Md. Saiful Islam, and Md. Ruhul Amin. 2022. Gca-net : Utilizing gated context attention for improving image forgery localization and detection. *Preprint*, arXiv:2112.04298.

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

Jingcheng Du, Peilin Jia, YuLin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. 2019. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82.

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

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2021. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *Preprint*, arXiv:2007.06225.

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

National Center for Biotechnology Information. 2024. Pubmed. `https://pubmed.ncbi.nlm.nih.gov`. Accessed: 2024-11-26.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. 2023. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures.

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

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *Preprint*, arXiv:2103.03206.

Ala Jararweh, Oladimeji Macaulay, David Arredondo, Olufunmilola M Oyebamiji, Yue Hu, Luis Tafoya, Yanfu Zhang, Kushal Virupakshappa, and Avinash Sahu. 2024. Litgene: a transformer-based model that uses contrastive learning to integrate textual information into gene representations. *bioRxiv*.

Xiaogang Jia, Songlei Jian, Yusong Tan, Yonggang Che, Wei Chen, and Zhengfa Liang. 2024. Gated cross-attention network for depth completion. *Preprint*, arXiv:2309.16301.

Alexander Pritzel Tim Green Michael Figurnov Olaf Ronneberger Kathryn Tunyasuvunakool Russ Bates Augustin Žídek Anna Potapenko1 Alex Bridgland Clemens Meyer Simon A. A. Kohl Andrew J. Ballard Andrew Cowie Bernardino Romera-Paredes Stanislav Nikolov Rishub Jain Jonas Adler Trevor Back Stig Petersen David Reiman Ellen Clancy Michal Zielinski Martin Steinegger Michalina Pacholska Tamas Berghammer Sebastian Bodenstein David Silver Oriol Vinyals Andrew W. Senior Koray Kavukcuoglu Pushmeet Kohli John Jumper, Richard Evans and Demis Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *nature*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. 2022. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Junnan Li, Dongxu Li, Chong Xiong, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. 2022b. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*.

Haotian Liu, Chunyuan Lin, Fangyun Zeng, and et al. 2023a. Llava: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*.

Leilei Liu, Xianglei Zhu, Yi Ma, Haiyin Piao, Yaodong Yang, Xiaotian Hao, Yue Fu, Li Wang, and Jiajie Peng. 2020. Combining sequence and network information to enhance protein–protein interaction prediction. *BMC Bioinformatics*, 21(16):537.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023b. Multimodal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024. Prott3: Protein-to-text generation for text-based protein understanding. *Preprint*, arXiv:2405.12564.

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

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

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. 2020. Progen: Language modeling for protein generation. *Preprint*, arXiv:2004.03497.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. 2022. Progen2: Exploring the boundaries of protein language models. *Preprint*, arXiv:2206.13517.

OpenAI et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sameer Quazi. 2022. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology*, 39(8):120.

Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. 2022. Whisper: Openai's speech recognition model. OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Guillaume Richard, Bernardo P. de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie Lopez, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, and Thomas Pierrot. 2024. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*.

Hartmut Schlüter, Rolf Apweiler, Hermann-Georg Holzhütter, and Peter R. Jungblut. 2009. Finding one's way in proteomics: a protein species nomenclature. *Chemistry Central Journal*, 3:11.

Gregory D Schuler, Jonathan A Epstein, Hitomi Ohkawa, and Jonathan A Kans. 1996. [10] entrez: Molecular biology database and retrieval system. In *Methods in enzymology*, volume 266, pages 141–162. Elsevier.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Attention-based wav2text with feature transfer learning. *Preprint*, arXiv:1709.07814.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Zhourun Wu, Mingyue Guo, Xiaopeng Jin, Junjie Chen, and Bin Liu. 2023b. CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics*, 39(3):btad123.

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

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. 2022. Ontoprotein: Protein pretraining with gene ontology embedding. *Preprint*, arXiv:2201.11147.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. 2023. A systematic study of joint representation learning on protein sequences and structures. *Preprint*, arXiv:2303.06275.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. 2024. Chemdfm: A large language foundation model for chemistry. *Preprint*, arXiv:2401.14818.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Dataset Collection

A.1 Pretraining Dataset

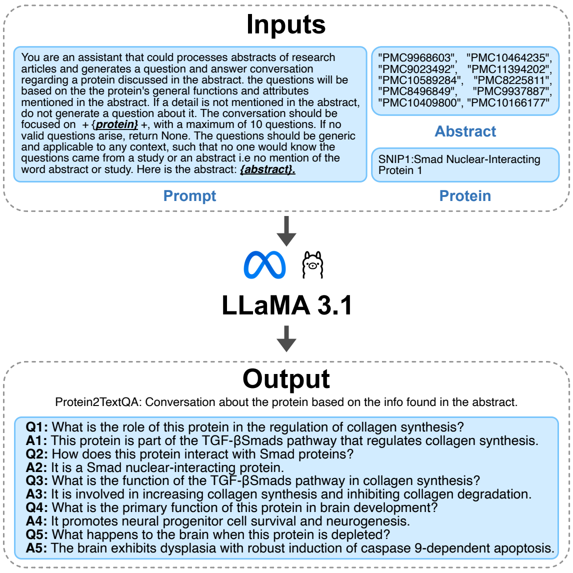
The pretraining dataset consists of protein sequences and their corresponding descriptive information obtained from UniProt [1] (Consortium, 2022). We have removed proteins with sequences that are longer than 2000 due to insufficient resources. These sequences when fed through the model, consume GPU VRAM and cause CUDA error even with a batch size of 1. We have not

---

[1] `https://www.uniprot.org`

Protein2Text-QA
Q1: What is the primary function of this protein in brain development? A1: It promotes neural progenitor cell survival and neurogenesis.
Q2: What happens to the brain when this protein is depleted? A2: The brain exhibits dysplasia with robust induction of caspase 9-dependent apoptosis.
Q3: How does this protein influence cell survival and death in the developing brain? A3: It regulates target genes that promote cell survival and neurogenesis.
Q4: What signaling pathways affect the activities of this protein? A4: TGF3b2 and NF3baB signaling pathways influence its activities.
Q5: What complex does this protein facilitate the genomic occupancy of? A5: It facilitates the genomic occupancy of Polycomb complex PRC2.
Q6: What is the general function of this protein? A6: This protein is involved in inhibiting transforming growth factor-3b2 (TGF-3b2) signaling, which is a process that helps regulate cell growth and division.

Table 7: **Sample of our Protein2Text-QA Data.** The data is extracted for the protein *"Smad nuclear-interacting protein 1"* with ID: *"Q8TAD8"*.



You are an assistant that could processes abstracts of research articles and generates a question and answer conversation regarding a protein discussed in the abstract. the questions will be based on the the protein's general functions and attributes mentioned in the abstract. If a detail is not mentioned in the abstract, do not generate a question about it. The conversation should be focused on + ***{protein}*** +, with a maximum of 10 questions. If no valid questions arise, return None. The questions should be generic and applicable to any context, such that no one would know the questions came from a study or an abstract i.e no mention of the word abstract or study. Here is the abstract: ***{abstract}.***

"PMC9968603", "PMC10464235", "PMC9023492", "PMC11394202", "PMC10589284", "PMC8225811", "PMC8496849", "PMC9937887", "PMC10409800", "PMC10166177"

SNIP1:Smad Nuclear-Interacting Protein 1

Protein2TextQA: Conversation about the protein based on the info found in the abstract.

Q1: What is the role of this protein in the regulation of collagen synthesis?
A1: This protein is part of the TGF-βSmads pathway that regulates collagen synthesis.
Q2: How does this protein interact with Smad proteins?
A2: It is a Smad nuclear-interacting protein.
Q3: What is the function of the TGF-βSmads pathway in collagen synthesis?
A3: It is involved in increasing collagen synthesis and inhibiting collagen degradation.
Q4: What is the primary function of this protein in brain development?
A4: It promotes neural progenitor cell survival and neurogenesis.
Q5: What happens to the brain when this protein is depleted?
A5: The brain exhibits dysplasia with robust induction of caspase 9-dependent apoptosis.

Figure 3: The pipeline to collect **Protein2Text-QA**. The prompt used to query the LLaMA3.1-Instruct model is comprised of three components: the role, the abstract extracted from PubMed (for Biotechnology Information, 2024), and the protein name to extract QA for.

performed any truncation to eliminate introducing noise to the model. We consider one specific prompt and its variant paraphrases such as *"Discuss the molecular function of this protein"*, *"Determine the function of this protein sequence"*, or *"Summarize the functional role of this protein sequence"*. The question, the function description (as the answer), and the protein sequence are used to create the dataset. Similar to image, instruction, and response in LLaVA (Liu et al., 2023a). This dataset is entirely used to train the resampler and the projector during the pretraining stage.

An example of the dataset is presented in Table 11, illustrating the structure and content of the data entries. Table 10 provides statistical details about the dataset, including the number of unique proteins, their variants, and the average lengths of sequences and descriptions. Variants—proteins derived from the same gene family—were carefully managed to ensure no data leakage, as all splits were performed based on unique proteins.

A.2 Finetuning Dataset: Protein2TextQA

The finetuning dataset (Protein2Text-QA) collection process involved two major steps: retrieving relevant abstracts from the literature and generating corresponding question-answer (QA) pairs using LLAMA3.

A.2.1 Abstract Retrieval

To collect protein-related abstracts, we used a systematic query approach with the PubMed Central (PMC) database (for Biotechnology Information, 2024). The queries targeted abstracts containing specific protein-related keywords. For each keyword, we performed a search using the `Entrez` library (Schuler et al., 1996), which interfaced with the PMC API. The search results returned lists of relevant PMC IDs, which were then used to fetch the abstracts. To ensure relevance, only abstracts explicitly mentioning the queried proteins were included.

Once retrieved, the abstracts were processed to remove redundant text (e.g., headings such as *Abstract*, *Methods*, and *Conclusion*) and cleaned of formatting inconsistencies. This preprocessing ensured that the text was suitable for input into the question generation pipeline.

A.2.2 Generating the QAs using LLaMA3

Figure 3 demonstrates the QA collection process pipeline. The cleaned abstracts, protein names, and the role were fed into **LLaMA3.1-8B-Instruct** (Dubey et al., 2024) to generate a conversation-style output. The model is prompted to generate a conversation between a chatbot and a human where the questions and answers are conditioned on the protein name mentioned in the prompt. The prompt instructed the LLaMA model to focus only on general protein functions and attributes explicitly mentioned while processing the abstract, ignoring other proteins. We limit the number of retrieved QA to up to 10 QA pairs per abstract.

The QA data are further preprocessed and tokenized to remove unnecessary questions that mention phrases such as *"no information found"*, *"answer not in the abstract"*, and *"not mentioned in the study"*. We attempt to make the questions general and related to the proteins instead of being related to the abstract. Table 7 shows a sample question and answers generated by LLaMA for the protein with ID *"Q8TAD8"*.

An example of the finetuning dataset is presented in Table 7, which highlights the structure of the QA pairs. The data extraction and question-generation pipeline, as implemented with LLaMA3 (Dubey et al., 2024), is demonstrated in Figure 3. The overall statistics of the finetuning dataset, including the number of QA pairs, unique proteins, and sequence lengths, are summarized in Table 10.

A.3 Evaluation Datasets

The evaluation datasets comprised four distinct subsets: Protein2Text QA test set, Zero-shot QA, DiscussionQA, and IntroductionQA. Each subset was curated to assess the model's performance.

First, the Protein2TextQA test set was randomly chosen from the entire Protein2TextQA without consideration of family or variant relationships. The protein sequences in the test set can be found in the pretraining dataset but not in the fine-tuning dataset.

Second, to generate the **Zero-shot QA** set, proteins and all their variants—defined as those from the same gene family—were entirely excluded from the training set. These proteins were included only in the test set, ensuring the model had no prior exposure to them during training. This dataset evaluates the model's ability to generalize to entirely unseen sequences.

The **Discussion QA** subset was derived using the same list of proteins from the test set subset. However, the QA pairs were generated from the *Discussion* sections of the corresponding research articles instead of the Abstracts. This subset tests the model's ability to handle context-specific questions derived from a different section of scientific texts. Similarly, the **Introduction QA** subset utilized the same list of proteins as the test set subset, but the QA pairs were generated from the *Introduction* sections of the articles. We were not able to extract introductions and abstracts for all of the articles, and we only considered proteins where we could find an introduction or discussion section that mentions them.

B Gated Cross Attention (GCA)

Gated-Cross Attention (GCA) (Alayrac et al., 2022; Jia et al., 2024; Das et al., 2022) attempts to find sampled media tokens that are influenced by the text tokens. For example, Alayrac et al. (2022) used GCA to allow the text modality to attend to the vision modality through a gating mechanism that controls the influence of the vision modality on text features. Here, we attempt to do the same approach but we allow the protein embeddings to attend to the text embeddings, aiming to create refined protein embeddings.

In our setup, the GCA operates after the projector. That is, it takes as input the projected protein embeddings ($\mathbf{H}'_v$) and the instruction embeddings ($\mathbf{H}_q$) and outputs text-informed protein embeddings ($\mathbf{H}''_v$).

The final text-informed protein embeddings ($\mathbf{H}'_v{}'$) are then concatenated to the original instruction embeddings and fed to the LLM decoder to generate the response.

$$\mathbf{H}''_v = \phi_{GCA}(\mathbf{H}'_v, \mathbf{H}_q)$$

$$\mathbf{H} = [\mathbf{H}''_v; \mathbf{H}_q]$$

Protein2Text	BLEU 2	BLEU 4	ROUGE 1	ROUGE 2	ROUGE L	METEOR	BERT Score	Biomed-BERT Score
-w/ resampler	**0.144**	**0.083**	**0.322**	**0.180**	**0.288**	**0.378**	**0.891**	**0.942**
-w/ GCA	0.1017	0.0596	0.3054	0.170	0.278	0.358	0.863	0.929

Table 8: Adding Gated-Cross Attention (GCA) on top of the resampler shows no performance improvement.

	Stage	Number of Trainable Parameters
Protein2Text *w/o resampler*	Pretraining Fine-tuning	22M 190M
Protein2Text *w/ Resampler*	Pretraining Fine-tuning	42M 232M
Protein2Text *w/ Resampler + GCA*	Pretraining Fine-tuning	76M 307M

Table 9: Number of parameters in various model architectures. Protein2Text w/o resampler refers to using only the projector (i.e. pure LLaVA model with changing the encoder).

The final set of tokens ($\mathbf{H}$) is fed to the LLM decoder to obtain the language response.

$$\text{response} = f_\phi(\mathbf{H}).$$

C Training

C.1 Training Procedure

The training consists of two main stages: pre-training and fine-tuning. Throughout the experiments in the manuscript, we use `LLaMA3.1-Instruct` model as the language decoder and `facebook/esm2_t33_650M_UR50D` as the protein encoder, unless otherwise specified. Every training stage is tailored to specific input, output, and training procedures. We now provide an overview of training details for every stage.

Pretraining. During pretraining, the model is expected to align the protein and the text modalities. Thus, we utilize protein sequences and their descriptions. During this stage only, the resampler and the projector are trained, aiming to learn the alignment between protein sequences and text. The dataset collected for this stage spans paraphrases on the question "Describe the function of the protein?". A sample of the dataset is shown in Table 11. We pre-train the model for one epoch following the LLaVA (Liu et al., 2023a) approach. The number of trainable parameters for the stage is 42 million (Table 9).

Finetuning. We next train the model to predict answers to a wide range of prompts where the prompt and the sequence are fed as input, and the response as the output. During this stage, the resampler, the projector, and the LLM are trained. We utilize LoRA (Low-Rank Adaptation) to train the model (Hu et al., 2021). LoRA freezes the pre-trained linear layers of the LLM architecture and learns a decomposition of two matrices of the frozen weights. The number of trainable parameters for this stage is 232 million parameters (190 million for LoRA adapters). The dataset used to train the model is a QA dataset. Refer to Appendix A and Table 7 for the dataset collection and an example conversation from the dataset respectively.

C.2 Hyperparameters

Since performing a parameter search to find the best-performing parameters is computationally intensive and exhaustive for LLMs (Benington et al., 2023), we rely on different factors to identify parameters. First, we inspect model parameters mentioned in previous studies in the same domain (Gu et al., 2021; Liu et al., 2024; Fang et al., 2024; Lin et al., 2022b) or similar domains (Liu et al., 2023a; Alayrac et al., 2022). Second, we track our training logs using Wandb to ensure the loss decreases for any respective ablation study.

Third, we also focus on targeted ablation studies to find the main parameters such as model sizes (i.e. ESM2-650 vs ESM2-3B). For example, Table 5 demonstrates reported ablation studies. We found that increasing model parameter

	Number of				Avg. Length		
Split	**QA Pairs**	**Sequences**	**Proteins**	**PMC IDs**	**Queries**	**Answers**	**Sequences**
Pretraining	393,849	393,849	70,854	0	8.8	42.1	378.4
Fine-tuning	209,847	5,556	5,574	29,198	12.3	12.9	511.0
Test QA	38,585	1044	1044	5,880	12.3	12.9	499.1
Zero-shot	14,107	348	348	2,164	12.2	13.0	433.9
DiscussionQA	2,629	180	180	263	12.9	17.3	385.6
IntroductionQA	1269	51	51	111	13.3	16.4	401.6

Table 10: Main statistics of the datasets used for the experiments in the study. Unique proteins can have different variants, and every variant has its sequence. These variants usually share a function similar to that of the dominant protein. We split based on unique proteins to eliminate data leakage. For the average length section, questions and answers are measured with words while sequences are measured in characters.

Protein ID/Name	**Description**
Q8TAD8: Smad nuclear-interacting protein 1	Required for pre-mRNA splicing as a component of the spliceosome. As a component of the minor spliceosome, involved in the splicing of U12-type introns in pre-mRNAs (Probable). Down-regulates NF-kappa-B signaling by competing with RELA for CREBBP/EP300 binding. Involved in the microRNA (miRNA) biogenesis. May be involved in cyclin-D1/CCND1 mRNA stability through the SNARP complex which associates with both the 3'end of the CCND1 gene and its mRNA.
Q8KAW9: ATP synthase gamma chain	Produces ATP from ADP in the presence of a proton gradient across the membrane. The gamma chain is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex.

Table 11: High overview of our pretraining data. The data is comprised of protein sequences and their descriptions from UniProt (Consortium, 2022).

size reduces the model performance and suggests the need for more data samples. Table 9 indicates the number of parameters for the main ablations performed. We found that increasing the number of latent tokens generated by the resampler from 128 to 256 worsened the performance of the model. Also, we saw adding gated cross-attention (Alayrac et al., 2022; Das et al., 2022; Jia et al., 2024) increases the number of parameters but decreases the performance. Refer to Section B for description about adding GCA, Table 9 for number number of parameters, and Table 8 for GCA results.

Model training and inferencing were mainly performed on 2 NVIDIA H100 PCIe GPUs of 80GB VRAM. The estimated training time is roughly dependent on the number of parameters, the batch sizes, and other configurations such as gradient checkpointing, LoRA parameters, and resampler configurations. However, the estimated training time for the pretraining stage varies from 8 to 13 hours while the fine-tuning stage varies from 12-20 hours. The time estimations are based on the parameters found in Table 9. The table also highlights the best-performing model parameters of the experiments in this manuscript.

D Expanded Discussion on Related Work

Instruction Tuning. Large Language Models (LLMs) have demonstrated significant capabilities in human understanding tasks, such as GPT models (Radford et al., 2019; Brown et al., 2020; OpenAI et al., 2023) and LLaMA models (Touvron et al., 2023a,b; Dubey et al., 2024). When LLMs were first introduced, they were mainly trained on next token prediction (Touvron et al., 2023a; Radford et al., 2019; Lewis et al., 2020; Liu et al., 2019; Yang et al., 2020). Instruction tuning has

Hyperparameter	**Pre-training**	**Fine-tuning**
Training		
Number of Epochs	1	5
Per-device Batch Size	10	5
Learning Rate	2×10^{-3}	8×10^{-6}
Max Sequence Length	2048 tokens	
Precision	bf16 (Mixed Precision)	
Optimizer	AdamW	
Gradient Accumulation Steps	1 step	
Warmup Ratio	0.03	
Protein Encoder		
Model	ESM2-650M	
Output Tokens	All (i.e. no truncation)	
Feature Layer	-2 (i.e. second to last)	
Language Model		
Model	LLaMA-3.1-8B-Instruct	
LoRA Rank	64	
LoRA Alpha	16	
Context Length	2048	
Projector		
Number of Layers	2 layers	
Activation	GELU	
Hidden Dimensions	4096	
Perceiver Resampler		
Number of Attention layers	4096	
Attention Heads	8	
Dimension of Attention Heads	4	
Multiplication Factor of Hidden State	2	
Number of Latent Tokens	128	

Table 12: An overview of the hyperparameters used to train the two stages of Protein2Text. If one parameter is mentioned across the two columns, the same value is used in the two training stages.

been proposed to align the training objective with the user objective by enhancing the model's ability to follow instructions (Zhang et al., 2024). Several models trained via instruction tuning have been proposed for a variety of tasks such as summarization (Basyal and Sanghvi, 2023), question answering (Ouyang et al., 2022; Muennighoff et al., 2023; Zheng et al., 2023), and zero-shot capabilities (Zheng et al., 2023; Ouyang et al., 2022; OpenAI et al., 2023; Wei et al., 2022; Dubey et al., 2024).

Multimodal LLMs. Multimodal LLMs have also been extensively applied to perform cross-modal tasks beyond the text modalities. For instance, several studies have been proposed to integrate vision and language (Liu et al., 2023a; Alayrac et al., 2022; Li et al., 2023), and integrate audio and language (Radford et al., 2022; Tjandra et al., 2017). Building on these efforts, LLMs have also witnessed prosperous adaptation to scientific and biomedical domains such as biomedical text understanding (Jararweh et al., 2024; Lee et al., 2019), biomedical QA (Wu et al., 2023a; Luo et al., 2023), clinical reasoning tasks (Huang et al., 2020), and molecular structure understanding (Zhao et al., 2024; Fang et al., 2024; Cao et al., 2023; Liu et al., 2023b).

Protein-related LLMs. The sequential nature of protein primary structure lends itself to language modeling for protein characterization. For

Model	**BLEU-2**	**BLEU-4**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**
ProtT3- *Fine-tuned*	**0.765**	**0.688**	**0.783**	**0.705**	**0.714**	**0.768**
Template predictor	0.243	0.219	0.667	0.621	0.667	0.498
Protein2Text- *Zero-shot*	0.277	0.217	0.447	0.345	0.383	0.396

Table 13: Performances on the ProteinKG25 (Zhang et al., 2022; Liu et al., 2024) benchmark. Template predictor refers to predicting the QA template as the response for all questions in the ProteinKG25 test set.

example, encoder-based LLMs trained on protein amino acid sequences have been adopted to generate a representation space that captures sequence structures (Lin et al., 2022b,a; Zhang et al., 2022; Elnaggar et al., 2021). Similarly, generative LLMs have also been proposed for a variety of protein generation tasks such as generating 3D structure (John Jumper and Hassabis, 2021), and novel protein sequences (Madani et al., 2020; Nijkamp et al., 2022; Lv et al., 2024). LLMs that incorporate natural language and protein as one modality (i.e. considering protein as text modality) have been proposed. For example, Galactica models are general-purpose LLMs that are trained on scientific corpora to perform different reasoning tasks including protein captioning. Leveraging advances in multimodal LLMs, several studies attempt to integrate text with protein modalities such as DNA/RNA sequences (Richard et al., 2024), 3D structure (Guo et al., 2023; Wang et al., 2024), and amino acid sequences (Xiao et al., 2024b; Luo et al., 2023). Similarly, multi-modality projection similar to vision-language alignment (Alayrac et al., 2022; Liu et al., 2023a), has been applied to align between protein and natural text where protein is considered as single modality (Guo et al., 2023; Wang et al., 2024; Liu et al., 2024; Luo et al., 2023; Fang et al., 2024).

E Baselines

We compare our model to different baselines throughout the manuscript. We mainly focus on two types of baselines: general-purpose LLMs and protein-specific LLMs. The general-purpose LLMs were used as a measure of data leakage, identifying the amount of information leaked from the prompt into the generated answer. Second, we assess protein-specific LLMs that use protein sequences and a text prompt as input. We now provide a high overview of the baselines and the prompting mechanism.

GPT4o-mini (OpenAI et al., 2023). The model is a variant of the GPT4 family with a reduced number of parameters. We used the OpenAI API to generate responses in this manuscript where we feed the prompt and the sequence as input. We set the role to "*You are an expert assistant for protein-related inquiries*". The average response time is 30 seconds per query. We launched multiple processes per day for multiple days until the maximum number of tokens quota was reached.

LLaMA3.1-8B-Instruct (Dubey et al., 2024). LLaMA3.1-8B-Instruct [2] is a general multilingual model trained using instruction tuning to perform reasoning tasks. We utilize the same prompt structure used to query GPT4o-mini to extract responses from the model. We use the released model checkpoints from HuggingFace to extract responses. The average request time is 30 seconds per prompt on an 80GB H100.

BioMedGPT (Luo et al., 2023). BioMedGPT is a multimodal LLM that integrates molecular structures, protein sequences, and natural language text. The model aligns the three modalities to perform cross-modal tasks about proteins and molecular compounds. The model utilizes LLaMA2 (Touvron et al., 2023b) as the LLM base model. The training data was extracted from different sources such as PubMed Central (PMC), PubChem (Kim et al., 2022), and UniProt (Consortium, 2022). We utilize the weights and default parameters released by the authors to perform inferencing. The inference time is 0.09 seconds per query on an 80GB H100.

Mol-Instruction (Fang et al., 2024). Similarly, Mol-Instruction is a multimodal LLM that integrates text, molecular compounds, and protein sequences. The model utilizes GPT3.5 to generate a QA dataset about proteins and compounds from PubMed articles. We utilize the LoRA weights published by the authors and the LLaMA-2-7b-chat-hf model from HuggingFace to perform inferencing. We utilize the default parameters as found

---

[2] `https://huggingface.co/blog/llama31?utm_source=chatgpt.com`

in the released evaluation script. The approximate inferencing time is 18.17 seconds per query on an 80GB A100.

ProtT3 (Liu et al., 2024). ProtT3 utilizes multimodal projection to align between protein amino acid sequences and natural language text. The model is trained in two stages: protein-text retrieval and protein-text generation. During the first stage, contrastive learning objectives are utilized to extract protein features that match the description. Then, the LLM model is trained using LoRA to perform generative tasks. The authors release three different checkpoints for different tasks. We utilize the checkpoint released by the author for the QA task. The response time is 0.14 seconds per query on an 80GB H100.

LitGene (Jararweh et al., 2024). LitGene is an encoder-based model that refines protein/gene embeddings by integrating textual descriptions and Gene Ontology (GO) terms. The model is designed for classification and retrieval tasks based on protein/gene embeddings. In this study, we use LitGene as a benchmark to evaluate our model ability in classification tasks. The results demonstrated in Figure 2 are based on benchmarks from the LitGene paper. We use their reported mean values on these benchmarks as a baseline for our model predictions.

F Benchmarks

ProteinKG25. The ProteinKG25 benchmark is a template-based dataset designed for protein captioning. The dataset is originally a gene ontology knowledge graph that consists of protein sequences, descriptions, and protein attributes (Zhang et al., 2022; Consortium et al., 2023). The authors of ProtT3 (Liu et al., 2024) synthesized a QA dataset based on the knowledge graph and used it for benchmarking. Table 14 shows a sample of the dataset, highlighting the template used to design the QA dataset from gene attributes.

Solubility. The solubility benchmark is a classification-based dataset that classifies whether a protein is soluble or insoluble. The dataset was collected by Jararweh et al. (2024) authors and used to benchmark their LLM-based model on the solubility task. The dataset originally consisted of protein descriptions and their respective classes. We further process the dataset into a QA format by extracting the sequences and adding the prompt *"Is this protein Soluble or Membrane?"*. The answer to this prompt would be the protein class: *"Soluble"* or *"Insoluble"*.

Localization. Similarly, the localization dataset is another classification benchmark from the LitGene paper Jararweh et al. (2024). The dataset is designed to classify the subcellular localization of proteins and spans the classes: *"Nucleus"*, *"Cytoplasm"*, and *"Cell Membrane"*. To design a QA dataset from this dataset, we extract protein sequences using the UniProt IDs and add the prompt: *"Is this protein localized in Nucleus, Cytoplasm, or Cell Membrane?"*. The answer to this prompt would be the protein label from the LitGene localization benchmark.

G Evaluation Metrics

BLEU Scores (Papineni et al., 2002). BLEU (Bilingual Evaluation Understudy) score relies on n-grams matching to calculate the performance of the generated text. The BLEU score is a precision-based metric that quantifies the number of n-grams in the generated text that are also mentioned in the ground-truth text. **BLEU-2** denotes bigrams matching while **BLEU-4** denotes 4-grams matching.

ROUGE Scores (Lin, 2004).Similarly, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score also focuses on sequence matching. As opposed to BLEU, the ROUGE score is recall-based which calculates the amount of n-grams from the ground truth that are captured by the generated response. **ROUGE-1** denotes unigram matching, **ROUGE-2** bigrams matching, and **ROUGE-L** denotes longest common subsequence matching.

METEOR Score (Banerjee and Lavie, 2005). METEOR (Metric for Evaluation of Translation with Explicit ORdering) weights recall and precision while performing n-gram matching. METEOR also captures high-level semantic similarity by applying stemming and synonym matching.

BERT Similarity Score (Devlin et al., 2019). We also utilize a similarity score generated using BERT. We embed the generated and ground truth sentences using the `bert-base-uncased` model, then we extract the embedding for the CLS token. The cosine similarity between the two sentences is then computed.

ProteinKG25
Q: What is the function of the protein? A: This protein is part of the following components: plasma membrane, membrane, and integral component of membrane. This protein is involved in the following processes: response to stimulus, signal transduction, sensory perception of taste, and G protein-coupled receptor signaling pathway. This protein is located in the following components: membrane, integral component of membrane, and plasma membrane. This protein is involved in signal transduction: signal transduction, and G protein-coupled receptor signaling pathway. This protein is part of membrane: membrane, and plasma membrane. This protein is part of integral component of membrane: integral component of membrane. This protein enables the following function: G protein-coupled receptor activity.

Table 14: ProteinKG25 QA (Zhang et al., 2022; Liu et al., 2024) is a QA benchmark built from a template of protein attributes.

BiomedBERT Similarity Score (Gu et al., 2021). We also utilize BiomedBERT, a variant of the BERT model but trained on text from the the biomedical domain. We use the CLS embedding of the generated and the ground truth texts to calculate the cosine similarity.

H Reproducibility and Availability

We release the model weights, evaluation scripts, and evaluation datasets at `https://github.com/alaaj27/Protein2Text.git`. Our efforts aim to enable results reproducibility and facilitate more rigorous evaluations of PLMs' generative capabilities across various tasks and different domains.

Protein ID/ Question	**Ground Truth**	**Model Outputs**
ID: P17041 **Q1**: Where does the protein localize?	The nucleus.	**Protein2Text:** The protein localizes to the nucleus. **LLaMA3:** The sequence you provided appears to be a long protein sequence, and it's difficult to pinpoint the localization without knowing the specific protein or its function. **ProtT3:** NUCLEAR PROTEIN **BioMedGPT:** Nucleus **Mol-Instructions:** Based on the input provided, the protein localizes to the endoplasmic reticulum.
ID: Q99489 **Q2**: What happens to the activity of this protein when it is nitrosylated?	Nitrosylation affects the activity of this protein to a limited extent.	**Protein2Text:** The activity of this protein is inhibited when it is nitrosylated. **LLaMA3:** Without specific information about the protein's function and the cysteine residues involved in nitrosylation, it's challenging to predict the exact effect of nitrosylation on protein activity. **ProtT3:** NO **BioMedGPT:** The activity of this protein is increased when it is nitrosylated. **Mol-Instructions:** Nitrosylation of this protein leads to a decrease in its enzymatic activity.
ID: G3V241 **Q3**: Is high-altitude environment characterized by an elevated abundance of this protein?	No, the abundance of this protein was decreased at high altitude.	**Protein2Text:** No, high-altitude environment is characterized by a lower abundance of this protein. **LLaMA3:** To determine if the high-altitude environment is characterized by an elevated abundance of the given protein, we would need to know more context about the protein and its function. **ProtT3:** MOLECULAR CLOCK PROTEIN, OXYGEN STORAGE, OXYGEN TRANSPORT **BioMedGPT:** Promotes the incorporation of iron into the crystal structure of the protein. **Mol-Instructions:** Yes, the high-altitude environment is characterized by an elevated abundance of the protein.
ID: C9JQQ2 **Q4**: Is this protein a canonical enzyme similar to some other proteins?	No, this protein is not a canonical enzyme similar to some other proteins.	**Protein2Text:** Yes, this protein is a canonical enzyme similar to some other proteins. **LLaMA3:** The protein sequence above is a member of the alpha/beta barrel fold family. It is a protein with a characteristic barrel-shaped structure, which is common among enzymes. **ProtT3:** NO **Mol-Instructions:** Based on the provided sequence, this protein appears to be a canonical enzyme similar to other proteins in the same family. The sequence shares high similarity with other proteins in the same family, indicating that it is likely to have similar structure and function. Additionally, the presence of conserved motifs and residues typical of enzymes in this family further supports this conclusion. Therefore, it is likely that this protein plays a similar role to other proteins in this family, such as catalyzing a specific chemical reaction.</s> **BioMedGPT:** This enzyme is a member of the ATP-grasp superfamily.

Table 15: Additional qualitative analysis of baseline models on protein-related questions.