ProteinGPT: Multimodal LLM for Protein Property Prediction and Structure Understanding

Yijia Xiao [1] **Edward Sun** [1] **Yiqiao Jin** [2] **Qifan Wang** [3] **Wei Wang** [1]

Abstract

Understanding biological processes, drug development, and biotechnological advancements requires a detailed analysis of protein structures and functions, a task that is inherently complex and time-consuming in traditional protein research. To streamline this process, we introduce ProteinGPT, a state-of-the-art multimodal large language model for proteins, which allows users to upload protein sequences and/or structures for comprehensive proteins analysis and responsive inquiries. ProteinGPT seamlessly integrates protein sequence and structure encoders with linear projection layers to ensure precise representation adaptation. It leverages a large language model (LLM) to generate accurate and contextually relevant responses. To train ProteinGPT, we construct a large-scale dataset of 132,092 proteins, each annotated with 20-30 property tags and 5-10 QA pairs per protein, and optimized the instruction-tuning process using GPT-4o. Experiments demonstrate that ProteinGPT effectively generates informative responses to protein-related questions, achieving high performance on both semantic and lexical metrics. It significantly outperforms baseline models and general-purpose LLMs in understanding and responding to protein-related queries. Our code and data are available at `https://github.com/ProteinGPT/ProteinGPT`.

1. Introduction

Proteins are fundamental molecular building blocks of life, playing critical roles in biological processes (Kitadai & Maruyama, 2018; Xiao et al., 2025). Understanding their structure, functions, and interactions is vital for advancements in drug discovery (Teague, 2003), healthcare (Organization & University, 2007), and biological/medical engineering (Kobsa & Saltzman). Recent breakthroughs in machine-learning-based protein structure and function prediction (Lin et al., 2023) have significantly accelerated biological research by reducing the reliance on traditional labor-intensive laboratory experiments and literature search.

Challenges. As proteins can be represented by strings of characters, each corresponding to an amino acid from an alphabet of 20 letters, recent advancements in Large Language Models (LLMs) have naturally extended to protein research. Existing protein LLMs such as ProtST (Xu et al., 2023), ProteinChat (Guo et al., 2023), and ProtChatGPT (Wang et al., 2024) focus primarily on sequence-based or structure-based modeling, limiting their ability to generate holistic protein insights from multiple modalities. For instance, protein sequences can reveal evolutionary information, functional sites, and sequence-structure relationships, while protein structures provide critical insights into spatial arrangement, structural dynamics, binding sites, and stability. Applying multimodal LLMs to protein modeling is non-trivial due to the challenge in aligning diverse modalities, such as textual descriptions, protein sequences, and protein structures. Meanwhile, direct end-to-end retraining for protein LLMs is usually impractical due to extensive requirements for annotated data.

Our Work. We propose ProteinGPT, a protein LLM that allows researchers to upload protein sequences and/or structures (via fasta or PDB files) and ask natural language questions. ProteinGPT consists of four major components: a protein sequence encoder, a protein structure encoder, a projection layer, and an LLM.

The *protein sequence encoder* is based on the ESM-2 (Evolutionary Scale Modeling 2) (Lin et al., 2023) model variant `esm2_t36_3B_UR50D` with 36 transformer layers and 3 billion parameters. Pretrained on `UniRef50/D` (Suzek et al., 2015), a comprehensive protein database that clusters sequences with at least 50% sequence identity and 80% coverage, this model ensures sequence diversity and informativeness in encoding. The *protein structure en-*

---

[1]University of California, Los Angeles, Los Angeles, CA [2]Georgia Institute of Technology, Atlanta, GA [3]Meta AI, Sunnyvale, CA. Correspondence to: Yijia Xiao <yijia.xiao@cs.ucla.edu>, Wei Wang <wei.wang@cs.ucla.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

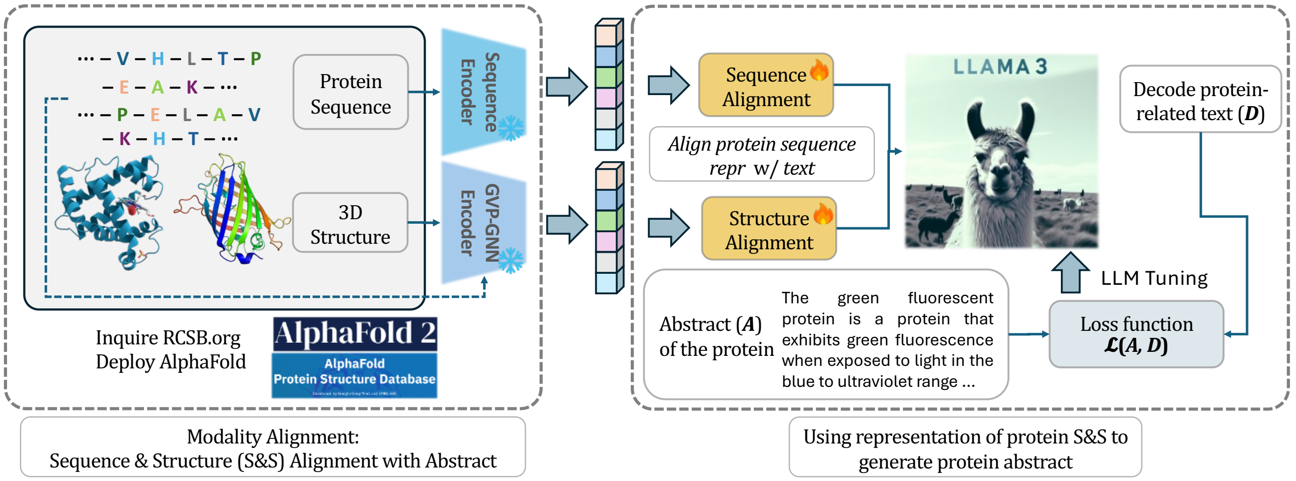

··· – V – H – L – T – P
– E – A – K – ···
··· – P – E – L – A – V
– K – H – T – ···
Protein Sequence
3D Structure
Sequence Encoder
GVP-GNN Encoder
Inquire RCSB.org
Deploy AlphaFold
AlphaFold 2
AlphaFold Protein Structure Database
Sequence Alignment
Align protein sequence repr w/ text
Structure Alignment
LLAMA 3
Decode protein-related text (D)
LLM Tuning
Abstract (A) of the protein
The green fluorescent protein is a protein that exhibits green fluorescence when exposed to light in the blue to ultraviolet range ...
Loss function $\mathcal{L}(A, D)$
Modality Alignment: Sequence & Structure (S&S) Alignment with Abstract
Using representation of protein S&S to generate protein abstract

Figure 1: ProteinGPT Modality Fusion & Alignment Stage: we freeze the encoder blocks and train the linear project layer to learn how to align protein structure and protein sequence representations with text. In the alignment stage, the input to the training is only the projected protein representation. No text prompts are incorporated in this stage.

coder, `esm_if1_gvp4_t16_142M_UR50`, is an inverse folding model that incorporates a geometric input processing layer paired with a seq2seq transformer (Bahdanau, 2015). Trained on 12 million structures predicted by AlphaFold2 (Jumper et al., 2021), the model effectively captures protein structural information.

To bridge these encoders with the LLM, we introduce a *projection layer* that aligns their embeddings with the LLM's latent space. This enables seamless integration of multimodal protein representations into the LLM, enabling information extraction from not only the protein structural and sequential information but also the rich pretrained knowledge of the ESM.

To train ProteinGPT for effective modality alignment, we introduce ProteinQA, a large-scale dataset of over 132,092 protein sequences with structural information and annotations. Unlike previous works that use entire protein annotations as prediction targets for instruction-tuning, we use GPT-4o (OpenAI et al., 2024) to systematically decompose proteins' abstract summary from the RCSB Protein Data Bank (RCSB PDB) (Burley et al., 2023) into question-answer (QA) pairs. We then finetune ProteinGPT on ProteinQA using diverse open-source models, including Llama-2 (Touvron et al., 2023), Llama-3 (Dubey et al., 2024) Vicuna (Chiang et al., 2023), and Mistral (Jiang et al., 2023). The training effectively enhances the model's ability to understand user queries and generate concise, contextually relevant answers. Our contributions are as follows:

- **Novel Framework.** We introduce ProteinGPT, a state-of-the-art protein LLM that fuses protein sequence and structural information to enable interactive protein-focused conversations, significantly enhancing the understanding and design of proteins;
- **Large-scale Dataset.** We propose ProteinQA, a large-scale protein dataset based on RCSB-PDB (Guo et al., 2023). ProteinQA encompasses 132,092 protein samples, each annotated with a detailed descriptive abstract, 20-30 property tags, and 5-10 QA pairs. The depth and variety of these annotations position ProteinQA as a high-quality instruction tuning corpus;
- **Comprehensive Experiments.** We conducted extensive experiments on mainstream open-source and proprietary LLM backbones under different scenarios. Our empirical analysis provides guidance for future design of protein LLMs.

2. Methodology

2.1. Model Architecture

ProteinGPT consists of two frozen pre-trained encoders (Figures 1 and 2): an inverse folding model (`esm_if1_gvp4_t16_142M_UR50`) for structure encoding and a protein language model for sequence encoding (`esm2_t36_3B_UR50D`). The embeddings generated by these models are fed into a linear projection layer to produce soft prompts for the LLM. The model training comprises two stages: 1) *Sequential and Structural Alignment* and 2) *Instruction-tuning*.

2.1.1. SEQUENCE AND STRUCTURE ALIGNMENT

In the alignment stage, protein structures are first fed into the pre-trained structure encoder `esm_if1_gvp4_t16_142M_UR50` which explicitly captures the detailed 3D structures and models

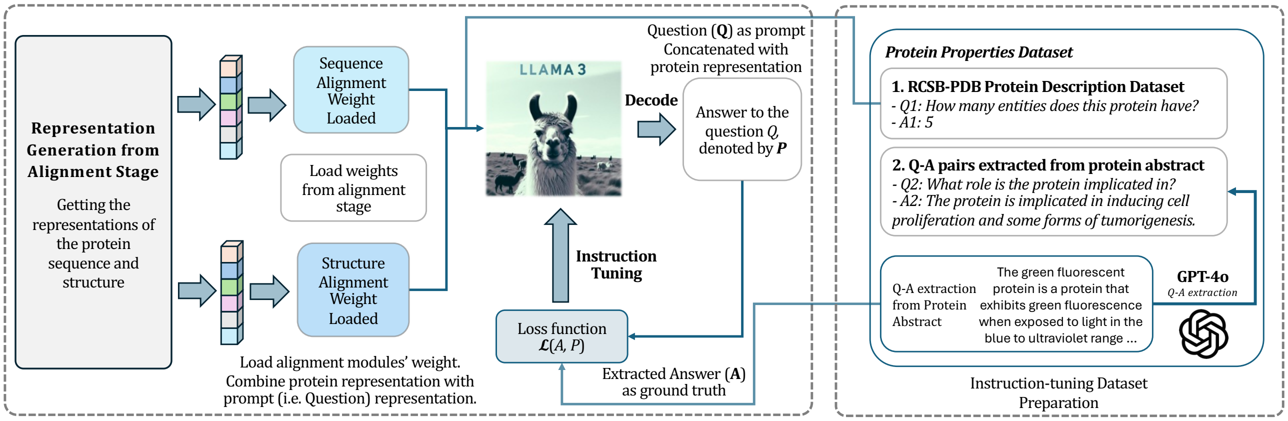

Representation Generation from Alignment Stage
Getting the representations of the protein sequence and structure
Sequence Alignment Weight Loaded
Load weights from alignment stage
Structure Alignment Weight Loaded
Load alignment modules' weight. Combine protein representation with prompt (i.e. Question) representation.
LLAMA 3
Question (Q) as prompt Concatenated with protein representation
Decode
Answer to the question Q, denoted by P
Instruction Tuning
Loss function ℒ(A, P)
Extracted Answer (A) as ground truth
Protein Properties Dataset
1. RCSB-PDB Protein Description Dataset
- Q1: How many entities does this protein have?
- A1: 5
2. Q-A pairs extracted from protein abstract
- Q2: What role is the protein implicated in?
- A2: The protein is implicated in inducing cell proliferation and some forms of tumorigenesis.
Q-A extraction from Protein Abstract
The green fluorescent protein is a protein that exhibits green fluorescence when exposed to light in the blue to ultraviolet range ...
GPT-4o
Q-A extraction
Instruction-tuning Dataset Preparation

Figure 2: ProteinGPT Instruction Tuning Stage: we utilize the QA pairs and property tags in ProteinQA to tune the LLM to follow instructions and give concise responses. For instruction alignment, explicit prompts (*Questions* on the protein) are included at the beginning of the prompt.

spatial interactions between amino acid residues. Then, sequences are encoded using the sequence encoder `esm2_t36_3B_UR50D` featuring 36 transformer layers and 3 billion parameters, trained on the Protein `UniRef50/D` database to enhance sequence diversity. This module integrates structural information with implicit structural contact, evolutionary, and biochemical information that the structure alone does not capture. For efficient training, both of these modules are frozen. We utilize a specialized token prompt for protein-text modality alignment:

$$\begin{aligned}&\mathbf{Q}: < Protein >< Struct >< Seq >\\&< /Protein >< QuestionPrompts >\\&\mathbf{A}: < Description >\end{aligned}$$

The structural and sequential information is encoded into the soft prompts and prepended to the question prompt. In stage 1 training, the question prompt **Q** is left empty to prioritize learning the abstract description from the protein representation.

The description tag is then replaced with the full annotation from RCSB-PDB (Guo et al., 2023) to train the projection layer in aligning a protein with its annotation description.

2.1.2. INSTRUCTION-TUNING

In stage 2, the model undergoes instruction tuning using our curated QA dataset. Unlike previous works that utilize full annotations, we focus on specific QA examples to facilitate instruction tuning. We augment the abstract dataset from stage 1 using GPT-4o to generate explicit QA pairs for this stage. The prompts from stage 1 are adapted to the LLaMA style ("`### Human:''` . . . and `### Assistant:` . . .), with **Q** replaced by explicit questions from the QA dataset, such as "how many assemblies does this protein have." The model then generates descriptive yet concise answers from the given protein as **A**.

2.2. Dataset Curation

To ensure the highest quality of training data, we implement a rigorous validation and data collection process. We leverage RCSB-PDB dataset (Berman, 2000), which is thoroughly vetted by both RCSB-PDB (Berman, 2000) and PubMed[1] for reliable protein information. The dataset is derived from peer-reviewed PubMed publications, implying that the dataset content is highly accurate. We select a large dataset to cover a wide range of proteins and ensures coverage on out-of-distribution proteins. For modality alignment, we construct a large-scale dataset from the RCSB-PDB database (Berman, 2000) consisting of 132,092 protein structures, sequences, and abstract descriptions. The raw dataset of 204,826 proteins is filtered to retain only those with an abstract description, chain A, and sequences without non-encodable characters. Each entry in the final dataset includes the 3D protein structure represented by backbone atomic coordinates, the sequence string, and a rich protein annotation, as shown in Figure 2. The detailed statistics of our dataset are presented in Table 1, highlighting the extensive annotations and comprehensive content available for each protein.

2.3. Data Augmentation

Previous works often use the entire protein annotation for instruction tuning (Xu et al., 2023; Guo et al., 2023; Wang et al., 2024), which may result in the model producing overly detailed responses with extraneous information not directly relevant to the user prompt. Therefore, our ProteinGPT decomposes the rich protein annotations into more specific QA-pairs for instruction tuning so that user instructions

---

[1] `https://pubmed.ncbi.nlm.nih.gov/`

can be concisely answered. We do this by generating short concise answers along with long-form responses to include our dataset. Concretely, we prompt GPT-4o to generate both open-ended and close-ended QA pairs with the context of the abstract to decompose the abstract into atom-level QA pairs. As seen in Table 1, on average, each protein has around 40 total QA pairs generated from this process.

3. Results

3.1. Quantitative Evaluation

3.1.1. EXPERIMENTAL SETUP

We evaluate our model using a curated dataset of 3,508 randomly-selected question-answer-protein pairs, covering 160 proteins from the test split. Each protein is associated with 28-30 questions that were not seen by ProteinGPT during training. We benchmark against several baseline models, including vanilla open-source LLMs (without modality alignment or instruction tuning) and proprietary models (GPT 4o/4/3.5, o3/o1-mini, and DeepSeek-R1). To manage computational costs, we sample 1,025 questions (35 proteins with 28$\sim$30 questions per protein). For fairness of comparison, we employ standard metrics to compare model predictions with the ground truth. These include 1) **semantic similarity** metrics that measure contextual meanings: BERTScore (S_{BERT}) (Zhang et al., 2020), PubMedBERT Score (S_{Pub}) (Gu et al., 2021), and GPT Score (S_{GPT}) (OpenAI, 2023); 2) **lexical quality** metrics that measure surface-level similarity based on n-gram overlaps: ROUGE-1/2/L (Ganesan, 2018). For the GPT Score, we use OpenAI's `text-embedding-3-large` as the embedding model.

3.1.2. COMPARISON AMONG PROTEINGPT VARIANTS

Among the 4 variants, ProteinGPT$_{\mathrm{Mistral}}$ performs the best in terms of both semantic and lexical metrics. In terms of BERTScore (Table 3), ProteinGPT$_{\mathrm{Mistral}}$ achieves 0.829, followed by ProteinGPT$_{\mathrm{LLaMA\text{-}3}}$ (0.790), ProteinGPT$_{\mathrm{LLaMA\text{-}2}}$ (0.764), and ProteinGPT$_{\mathrm{Vicuna}}$ (0.756). The strong performance of ProteinGPT$_{\mathrm{Mistral}}$ can be attributed to its integration of sliding window attention (SWA) (Jiang et al., 2023). Protein sequences are inherently lengthy and complex, often requiring models to capture intricate dependencies across extended stretches of amino acid. SWA helps capture local patterns and dependencies within protein sequences, leading to a longer effective attention span crucial for tasks like secondary structure prediction and functional annotation. The reduced computational load associated with SWA allows for the processing of longer protein sequences without a proportional increase in resource consumption. Figure 5 shows the visual comparison among different base LLMs of ProteinGPT.

3.1.3. BASELINE COMPARISON

We also compare ProteinGPT to three groups of baselines to demonstrate ProteinGPT's effectiveness in protein-specific multimodal tasks: 1) *Vanilla Open-source LLMs* (Vicuna, Mistral, LLaMA-3, and LLaMA-2); 2) *Proprietary General-Purpose LLMs* (GPT-4o/4/3.5); 3) *State-of-the-art models with Strong Reasoning Capabilities* (OpenAI o1/o3-mini and DeepSeek-R1). For fairness of comparison, we prepended the protein's FASTA sequence to the prompt to provide context for these LLMs. Tables 3 and 5 show model performance with semantic and lexical scores.

Vanilla open-source LLMs exhibit low semantic performance. When providing protein sequences as text input (Table 3a), the BERTScore (S_{BERT}) range from 0.490 (LLaMA-3) to 0.572 (Vicuna) in terms of precision, indicating a lack of protein-specific pretraining and limited semantic understanding of protein data. Meanwhile, incorporating modality fusion (where additional cues beyond the raw protein sequence are integrated) leads to modest improvements, such as an improvement from 0.572 to 0.582 when using Vicuna as the base model. Model Similarly, **proprietary models** such as GPT-4o/4/3.5, OpenAI o1/o3-mini, and DeepSeek-R1 also exhibit lower semantic and lexical performance when processing protein sequences as text inputs. This performance gap is likely due to these models not being pretrained on domain-specific data.

The results in Table 4 highlight a critical challenge in applying general-purpose language models to protein-related queries: standard models, even those with strong reasoning capabilities like o3-mini and DeepSeek-R1, struggle to interpret protein sequences. o3-mini achieves a marginally higher Rouge-L score (0.072) than GPT-4o (0.067), but remains far below ProteinGPT$_{\mathrm{Mistral}}$ (0.460). DeepSeek-R1 performs particularly poorly, which suggests that it struggles to extract meaningful insights from protein sequences when they are formatted as text. The sample answers in Appendix G.3 show that, while models like DeepSeek-R1 demonstrates strong reasoning capabilities on general QA tasks, it struggles with domain-specific terminology such as `assemblies`, interpreting inputs merely as a `long string of amino acids`. It tends to generate verbose, speculative responses, often introducing uncertainty by stating `Maybe the question is about...`–which diminishes its utility for precise scientific queries. As a result, its performance is comparable to or slightly worse than the modality-aligned version of ProteinGPT.

The overall comparison can be seen in Figure 3, which shows that ProteinGPT outperforms both baselines consistently. This demonstrates that our model outperforms knowledge embedded within LLMs and effectively utilizes sequence and structure information to answer questions.

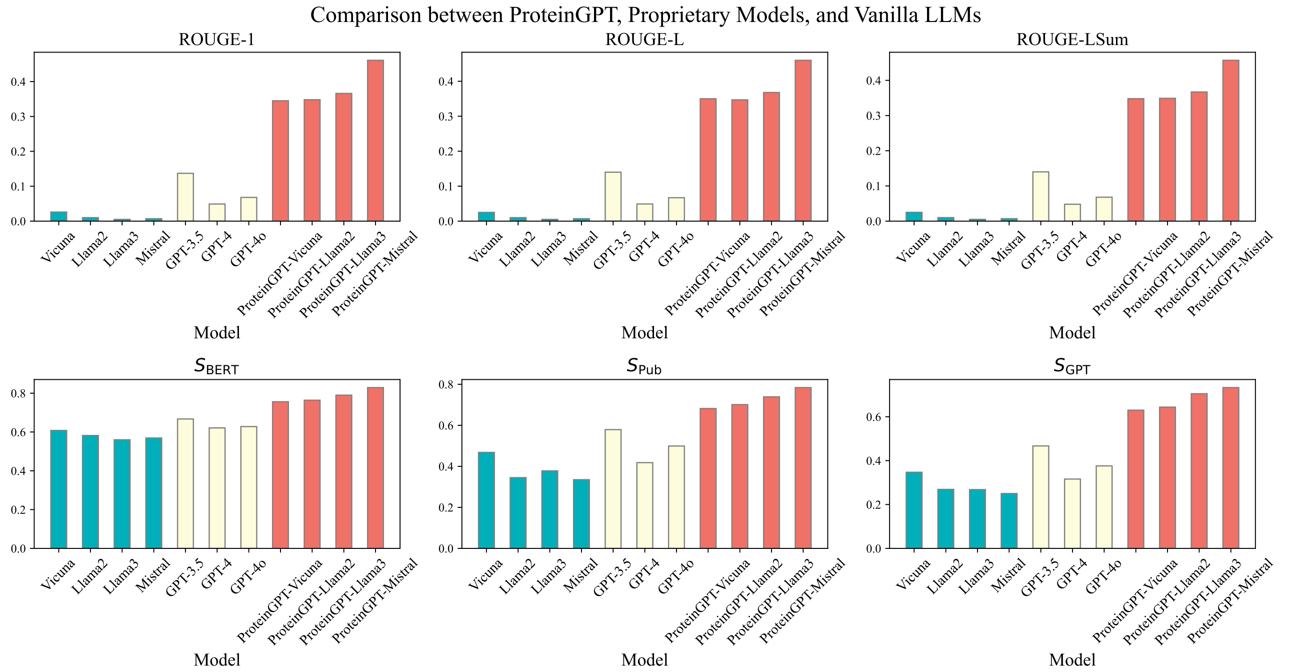


Figure 3: Protein Text LLM takes protein primary sequence as part of the prompt to the model. GPT models are more powerful than open-source LLMs like LLaMA and Mistral. Given the same protein sequence as input, ProteinGPT utilizes the information from sequence and structure encoders and yields more accurate responses.

Baseline Comparisons to ProteinGPT We also compare ProteinGPT to two baselines to demonstrate our contributions in creating multimodal LLMs that are more capable than general-purpose LLMs in communicating about proteins. The first baseline is the vanilla LLMs that we trained our models on, such as Vicuna, Mistral, LLaMA-3, and LLaMA-2. The second baseline is GPT-4 and GPT-3.5. For evaluation, we simply pretended the FASTA sequence of the protein in front of the prompt to give the LLM context of the protein. Table 2 and Figure 3 provide quantitative and visual comparison on baseline LLMs with ProteinGPT.

3.1.4. CLOSE-ENDED ACCURACY EXPERIMENT

Although semantic-based evaluations may be useful in gauging the feasibility of our outputs, to ensure our model is outputting factually correct information regarding a given protein, we also conduct a close-ended answer format evaluation on ProteinGPT with samples from our test subset of proteins. We selected 160 proteins for evaluation but only used QA-pairs that had a factual single-word ground truth and excluded questions that had open-ended answers (e.g. ”describe this protein”). Examples of such closed-ended questions are ”yes”/”no” questions or information on the number of assemblies or polymers in a protein. We then use GPT-4o to directly judge the outputs of ProteinGPT to the ground truth in our dataset.

The results can be seen in Figure 6. LLaMA-3 and Mistral are the best-performing backbone models, achieving around 80% accuracy in answering fact-based closed-ended questions. Even the weaker models like LLaMA-2 and Vicuna achieve reasonable accuracy above 70%. Therefore, ProteinGPT not only demonstrates strong capabilities in generating feasible answers as demonstrated by our semantic evaluations, but ProteinGPT also provides factually accurate answers as demonstrated by this accuracy evaluation.

4. Conclusions

We introduce ProteinGPT, a protein LLM that enhances question-answering capabilities and facilitates protein understanding with concise, informative responses. ProteinGPT fuses structure with sequence modalities and enables alignments to any base LLMs. Our results demonstrate ProteinGPT ’s potential for practical applications in protein understanding and design, highlighting the value of interactive protein models as dynamic research tools. Looking ahead, future enhancements aim to introduce multi-user support, enabling real-time collaboration and knowledge sharing in biological research. Additionally, we are developing user-friendly interfaces and integration with existing lab workflows, ensuring effortless adoption into bioinformatics and computational biology pipelines. By bridging ProteinGPT with widely used tools, we aim to drive innovation and collaboration in protein research.

References

Bahdanau, D. Neural machine translation by jointly learning to align and translate. In *arXiv arXiv:1409.0473*, 2015.

Berman, H. M. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. ISSN 1362-4962. doi: 10.1093/nar/28.1.235. URL `http://dx.doi.org/10.1093/nar/28.1.235`.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., Craig, P. A., Crichlow, G. V., Dalenberg, K., Duarte, J. M., et al. Rcsb protein data bank (rcsb. org): delivery of experimentally-determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic acids research*, 51(D1):D488–D508, 2023.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

Frohberg, J. and Binder, F. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models, 2022.

Ganesan, K. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*, 2018.

Gligorijevi'c, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021. ISSN 2637-8051. doi: 10.1145/3458754. URL `http://dx.doi.org/10.1145/3458754`.

Guo, H., Huo, M., and Xie, P. Proteinchat: Towards enabling chatgpt-like capabilities on protein 3d structures. 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv:2310.06825*, 2023.

Jin, W., Wohlwend, J., Barzilay, R., and Jaakkola, T. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

Kitadai, N. and Maruyama, S. Origins of building blocks of life: A review. *Geoscience Frontiers*, 9(4):1117–1153, 2018.

Kobsa, S. and Saltzman, W. M. Bioengineering approaches to controlled protein delivery. URL `https://www.nature.com/articles/pr2008103`.

Kong, X., Huang, W., and Liu, Y. Conditional antibody design as 3d equivariant graph translation. *arXiv preprint arXiv:2208.06073*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019.

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.

Mahbub, S. and Bayzid, M. S. Egret: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Briefings in Bioinformatics*, 23(2):bbab578, 2022.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed precision training, 2018.

Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., and Yuan, L. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023.

Notin, P., Dias, M., Frazer, J., Hurtado, J. M., Gomez, A. N., Marks, D., and Gal, Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017, 2022.

OpenAI. Gpt-4 technical report. *Arxiv Preprint*, arXiv:2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

Organization, W. H. and University, U. N. *Protein and amino acid requirements in human nutrition*, volume 935. World Health Organization, 2007.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pp. 2020–12, 2020.

R'eau, M., Renaud, N., Xue, L. C., and Bonvin, A. M. Deeprank-gnn: a graph neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*, 39(1):btac759, 2023.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Shu, F., Zhang, L., Jiang, H., and Xie, C. Audio-visual llm for video understanding, 2023.

Suzek, B., Wang, Y., Huang, H., McGarvey, P., and Wu, C. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 2015.

Teague, S. J. Implications of protein flexibility for drug discovery. *Nature Reviews Drug Discovery*, 2(7):527–541, July 2003. ISSN 1474-1784. doi: 10.1038/nrd1129. URL http://dx.doi.org/10.1038/nrd1129.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

Wang, C., Fan, H., Quan, R., and Yang, Y. Protchatgpt: Towards understanding proteins with large language models, 2024.

Xiao, Y., Zhao, W., Zhang, J., Jin, Y., Zhang, H., Ren, Z., Sun, R., Wang, H., Wan, G., Lu, P., et al. Protein large language models: A comprehensive survey. *arXiv:2502.17504*, 2025.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multi-modality learning of protein sequences and biomedical texts, 2023.

Yao, B., Jiang, M., Yang, D., and Hu, J. Benchmarking llm-based machine translation on cultural awareness, 2024.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.

Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., and Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023a.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. Benchmarking large language models for news summarization, 2023b.

Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations*, 2023c.

Per Protein	Min	Max	Mean
# Abstracts tokens	89	728	205.45
# Answer tokens	32	550	98.56
# Open-Ended QA Pair samples	10	26	17.39
# Closed-Ended QA Pair samples	24	29	25.94

Table 1: Statistics for ProteinQA

A. Experiments

A.1. Training

We trained ProteinGPT on 4 base LLM architectures: Vicuna (Chiang et al., 2023), LLaMA-2 (Zhang et al., 2023a), LLaMA-3 (Dubey et al., 2024), and Mistral (Jiang et al., 2023). Our training process is divided into two phases: modality alignment (MA) and instruction tuning (IT). This approach allows the model to preserve previously acquired knowledge while effectively handling specific instructions, such as protein-related queries.

Stage I: Modality Fusion/Alignment (MA). In this stage, we focus exclusively on training the projection adapter by freezing both sequence and structure encoders. We set the maximum text length of abstracts to 384 characters to accommodate the annotation lengths within the RCSB-PDB dataset. The projection layer is trained over 10 epochs with a batch size of 1, weight decay of 0.05, and 2048 warm-up steps. The dataset is divided into a training set (70%) of 105,673 proteins and a testing set (30%) of 26,419 proteins. We utilize the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$ (Loshchilov & Hutter, 2019), and employ a learning rate scheduler with a linear warm-up followed by cosine annealing. We set the initial learning rate to 1×10^{-4}, the minimum learning rate to 8×10^{-5}, and the warm-up learning rate to 1×10^{-6}. Automatic mixed precision (AMP) (Micikevicius et al., 2018) was used to improve training efficiency.

Stage II: Instruction Tuning (IT). In this stage, the model is fine-tuned on a protein question-answering task. Training is conducted for 10 epochs with a batch size of 1, weight decay of 0.05, and 200 warm-up steps. The QA dataset used in this phase comprises approximately 3.7 million samples, with around 35 questions per protein. We apply similar settings for the AdamW optimizer and AMP, but with a lower initial learning rate of 1×10^{-5}, a minimum rate of 1×10^{-6}, and a warm-up rate of 1×10^{-6}.

A.2. Inference

In real-world scenarios, there are cases where only protein sequence or structure information is available. Under such scenarios, protein folding and protein inverse-folding models are applied to obtain the missing structure or sequence information.

For **missing or incomplete sequences**, we use the `ESM-IF1` inverse folding model to predict absent segments, generating protein sequences from backbone atom coordinates. This method is effective in filling sequence gaps where traditional approaches struggle.

For **missing structural data**, we apply `ESMFold` to predict 3D structures of missing regions, generating accurate 3D conformations without relying on multiple sequence alignments.

A.3. Computational Cost

Our training uses two NVIDIA H100 PCIe GPUs (80GB vRAM) and two NVIDIA A100 PCIe GPUs (40GB). We implemented strategies including Automatic mixed precision (AMP), optimized data loaders, asynchronous data processing, multi-GPU training. Stage 1 and 2 require approximately one week and 60 hours, respectively.

ProteinGPT is optimized for fast processing, handling user protein queries in ~ 50 seconds and generating 23-token responses in about 8 seconds. Its computational efficiency allows deployment on hardware ranging from GPUs like the NVIDIA T4 with 16 GB memory to standard CPUs.

B. Results

We conducted a series of experiments to assess ProteinGPT's effectiveness both quantitatively and qualitatively. Moreover, we benchmark the capabilities of state-of-the-art large language models on protein-related tasks.

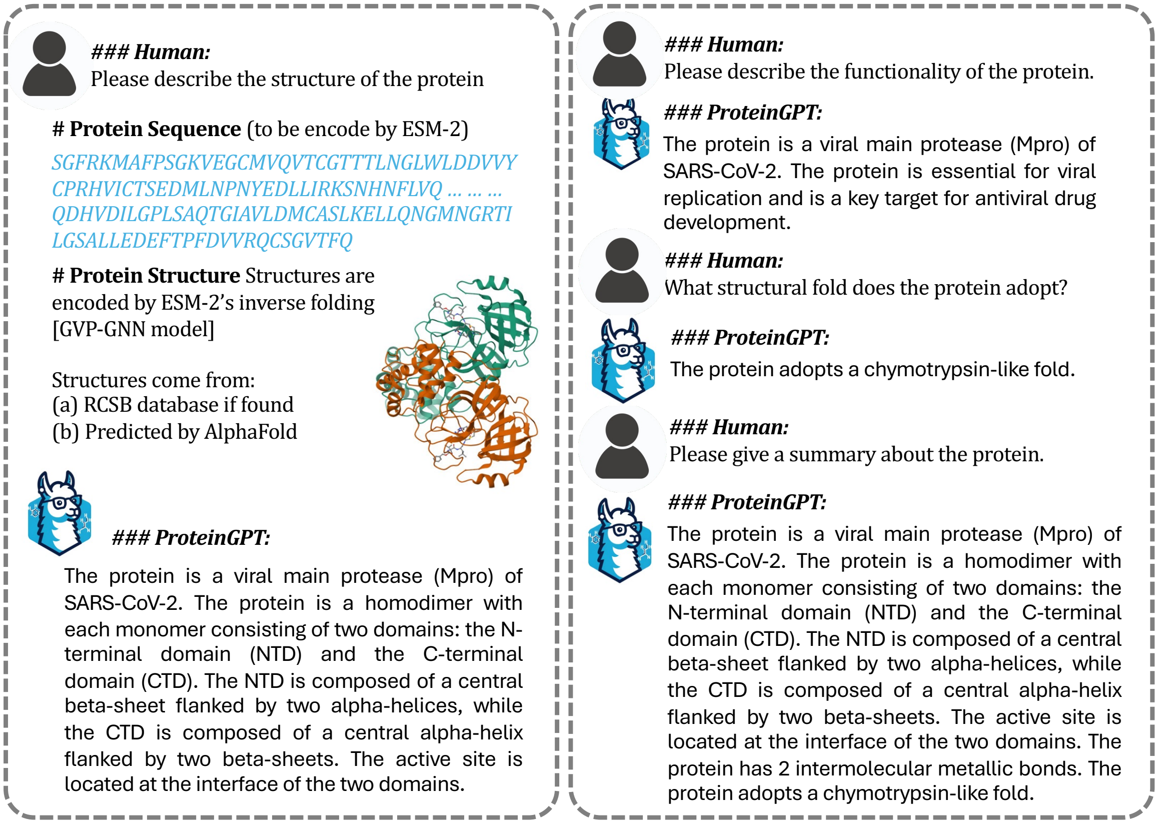
B.1. Qualitative Evaluation

Figures 4, 9, 11 show example conversations between human users and ProteinGPT on Proteins 6O7Q, 5X1Y, and 7RUV, respectively. To ensure unbiased evaluation and avoid data leakage, all testing was performed on a separate set of proteins isolated from training. ProteinGPT effectively interpret the semantics of queries and produces accurate, logically consistent responses. These responses include details about protein *functions*, such as catalyzing the reduction of dinitrogen to ammonia, and *structures*, such as the structural dependencies on the substrate azide and the product ammonia. This demonstrates ProteinGPT's capabilities on protein sequence, structure, and function understanding tasks and its potential for enabling rapid exploration of proteins.

B.2. Ablation Study

Lastly, to better understand our contributions and ensure the model is learning at each step in the architecture, we perform an ablation study to demonstrate that the module after each stage is indeed improving the performance of ProteinGPT. ProteinGPT is broken down into three modules: vanilla LLM, modality-aligned LLM, and instruction-tuned LLM.

Vanilla LLM: Table 3 (a) and Table 5 (a) display the seman-



### Human:
Please describe the structure of the protein

# Protein Sequence (to be encode by ESM-2)

SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVY
CPRHVICTSEDMLNPNYEDLLIRKSNHNFLVQ ... ...
QDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTI
LGSALLEDEFTPFDVVRQCSGVTFQ

# Protein Structure Structures are encoded by ESM-2's inverse folding [GVP-GNN model]

Structures come from:
(a) RCSB database if found
(b) Predicted by AlphaFold

### ProteinGPT:

The protein is a viral main protease (Mpro) of SARS-CoV-2. The protein is a homodimer with each monomer consisting of two domains: the N-terminal domain (NTD) and the C-terminal domain (CTD). The NTD is composed of a central beta-sheet flanked by two alpha-helices, while the CTD is composed of a central alpha-helix flanked by two beta-sheets. The active site is located at the interface of the two domains.

### Human:
Please describe the functionality of the protein.

### ProteinGPT:
The protein is a viral main protease (Mpro) of SARS-CoV-2. The protein is essential for viral replication and is a key target for antiviral drug development.

### Human:
What structural fold does the protein adopt?

### ProteinGPT:
The protein adopts a chymotrypsin-like fold.

### Human:
Please give a summary about the protein.

### ProteinGPT:
The protein is a viral main protease (Mpro) of SARS-CoV-2. The protein is a homodimer with each monomer consisting of two domains: the N-terminal domain (NTD) and the C-terminal domain (CTD). The NTD is composed of a central beta-sheet flanked by two alpha-helices, while the CTD is composed of a central alpha-helix flanked by two beta-sheets. The active site is located at the interface of the two domains. The protein has 2 intermolecular metallic bonds. The protein adopts a chymotrypsin-like fold.

Figure 4: Conversations between humans and ProteinGPT on Protein 6O7Q, where ProteinGPT provides detailed insights into both sequence (e.g., *60-subunit MoFe proteins*) and structural information (e.g., *substrate azide and product ammonia*).

Scoring Metrics		GPT-3.5	GPT-4	GPT-4o	P-GPT$_{\text{Vicuna}}$	P-GPT$_{\text{Llama-2}}$	P-GPT$_{\text{Llama-3}}$	P-GPT$_{\text{Mistral}}$	o1-mini	o3-mini	DeepSeek-r1
$\mathbf{S}_{\text{BERT}}$	Precision	0.641	0.578	0.596	0.730	0.739	0.779	**0.821**	0.572	*0.597*	0.468
	Recall	0.701	0.675	0.668	0.788	0.796	0.803	**0.839**	0.664	*0.699*	0.635
	F1	0.667	0.621	0.628	0.756	0.764	0.790	**0.829**	0.612	*0.641*	0.537
$\mathbf{S}_{\text{Pub}}$	Precision	0.513	0.434	0.440	0.626	0.644	0.706	**0.758**	0.402	*0.450*	0.449
	Recall	0.667	0.406	0.580	0.751	0.773	0.776	**0.816**	0.393	*0.513*	0.307
	F1	0.579	0.418	0.499	0.682	0.701	0.739	**0.784**	0.397	*0.477*	0.363
$\mathbf{S}_{\text{GPT}}$	Precision	0.470	0.391	0.391	0.605	0.606	0.698	**0.717**	0.387	*0.410*	0.360
	Recall	0.466	0.266	0.363	0.661	0.689	0.713	**0.752**	0.281	*0.396*	0.247
	F1	0.467	0.316	0.376	0.630	0.644	0.705	**0.733**	0.325	*0.398*	0.293

Table 2: Semantic Performance on OpenAI GPT Protein Text LLMs (GPT-3.5, GPT-4, and GPT-4o) and our ProteinGPT$_{\{\text{Vicuna, Llama-2, Llama-3, Mistral}\}}$ models in terms of BERTScore (S_{BERT}) (Zhang et al., 2020), PubMedBERT Score (S_{Pub}) (Gu et al., 2021), and GPT-4o Score (S_{GPT}) (OpenAI et al., 2024).

tic and lexical scoring using similar metrics for the vanilla LLM of choice (Vicuna, LLaMA, etc.). This is evaluated on the same set of 160 proteins and 3508 questions that we used to evaluate the final model. Also note that at this stage, no training has been done, the model is the same as the out-of-the-box LLM.

Modality Aligned (MA) LLM: Following this, Table 3 (b) and Table 5 (b) show that of the LLM after modality alignment. Evaluated on the same set of proteins, at this stage, the linear layer has been trained to learn to align and fuse the structure and sequence modalities to the LLM.

Instruction Tuned (IT) LLM: Lastly, as mentioned previously, Tables 3 (c) and 5 (c) are for the fully aligned and instruction-tuned model. At this stage, the model is complete and has been tuned on our GPT-4o curated dataset to follow instructions concisely.

Figure 7 highlights the differences in performance between each of these stages. We can observe that consistently, the instruction-tuned and modality-aligned final model outper-

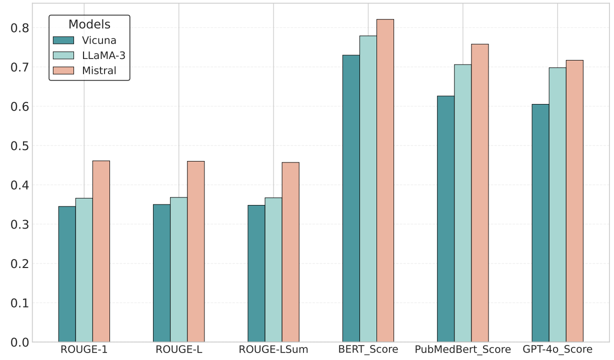


Figure 5: ProteinGPT performance with various base LLMs.

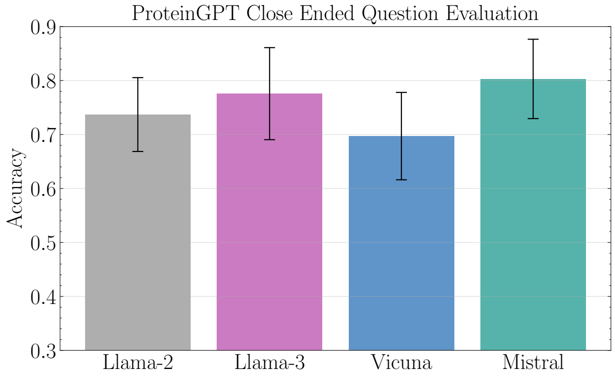


Figure 6: ProteinGPT performance (after instruction-tuning) on fact-based, closed-ended questions, such as determining number of polymer entities in a given protein.

forms the modality-only model and vanilla LLMs. This falls in line with our hypothesis and demonstrates that our 3 stages of training are indeed improving the model's multimodal understanding of proteins. More specifically, the observation that modality alignment always performs better than a vanilla LLM demonstrates that through this stage, the MLLM understands how to digest multi-modal information. Similarly, the observation that the instruction-tuned and modality-aligned ProteinGPT performs better than all other stages demonstrates that this stage indeed teaches the model how to properly answer questions related to these structures and sequences it learned from the previous stage.

C. Limitations

While ProteinGPT demonstrates strong capabilities in protein sequence and structure understanding, there are areas that can be improved: 1) *Potential for Hallucination.* As with most LLM-based systems, ProteinGPT may occasionally generate responses that are not fully aligned with established biological knowledge. However, its integration of protein-specific encoders helps mitigate this risk by grounding predictions in structured data. Future work can further refine this by incorporating confidence scores or uncertainty estimation mechanisms. 2) *Verifiability.* Currently, ProteinGPT does not provide direct citations for its responses, which may make it challenging to trace specific claims back to primary sources. We did not include citation data into the current training set of ProteinGPT due to the scarcity of protein datasets with reliable, consistent, and fine-grained citations that link specific claims or answers to appropriate references. As a result, integrating this data could potentially lead to inaccurate answers. Techniques such as retrieval augmented generation (RAG) () or explicit literature grounding for better reliability and trustworthiness. 3) *Training Data*. As with any data-driven model, the performance of ProteinGPT is influenced by the quality and diversity of the training data. While we carefully curated a large-scale dataset of 132,092 proteins with structured annotations, different processing & alignment strategies, continuous enhancement, and feedback from biological experts can further optimize its performance.

As demonstrated with alignment using GPT-4, different strategies may yield vastly different results.

D. Open Access and Deployment

To maximize accessibility, usability, and reproducibility, we will open-source both ProteinGPT and the ProteinQA dataset, allowing researchers to experiment with different backbone LLMs and protein encoders. ProteinGPT's flexible design ensures adaptability to model architectures.

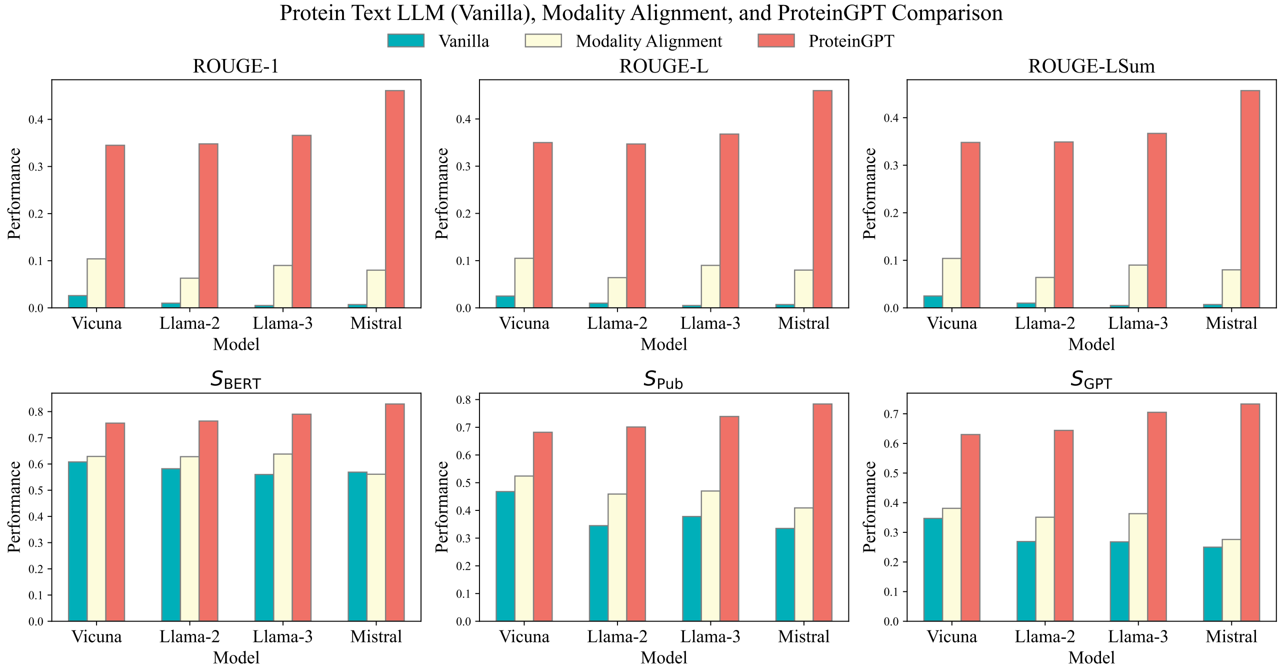


Figure 7: Performance improves progressively from the vanilla LLM model with protein as text to the modality-aligned version, and finally to the instruction-tuned variants of ProteinGPT. Each stage of ProteinGPT's training results in substantial enhancements in both lexical and semantic performance, showcasing the efficiency of our framework.

E. Additional Experiments

E.1. Case Studies

To avoid data leakage, we reserved 30% of our QA and abstract dataset for testing, which is around 26,419 proteins. This ensures that the tests reflect real-world scenarios as ProteinGPT has never seen these proteins before during training. We provide ProteinGPT's Q and A on Protein 7RUV in Figure 11.

F. Comparative Plots

Figures 5 and 10 highlight the performance comparison across different models, including various versions of ProteinGPT, OpenAI's GPT models, and Mistral models, using multiple evaluation metrics.

Figure 10 shows the performance of ProteinGPT variants alongside OpenAI's GPT and MA models. The key observations from this figure are:

- **ROUGE-1 and ROUGE-L**: The ProteinGPT_Mistral model significantly outperforms the other models with scores of 0.451 and 0.460, respectively. ProteinGPT_LLaMA-3 and the other GPT variants (GPT-3.5-turbo, GPT-4-turbo, GPT-4o) have much lower scores, indicating that the Mistral-based variant is superior in these metrics.
- **ROUGE-LSum**: Similar to ROUGE-1 and ROUGE-L, the ProteinGPT_Mistral variant leads with a score of 0.457, followed closely by ProteinGPT_LLaMA-3 at 0.387. Other models show significantly lower scores, emphasizing the effectiveness of the Mistral variant.
- **BERT Score** (S_{BERT}): The ProteinGPT_Mistral model also performs best with a score of 0.821, with ProteinGPT_LLaMA-3 following at 0.779. The GPT models lag behind, demonstrating that the ProteinGPT variants are more aligned with human evaluations.
- **PubMedBert Score** (S_{Pub}): Again, ProteinGPT_Mistral achieves the highest score of 0.758, outperforming ProteinGPT_LLaMA-3 slightly. The GPT models perform lower in this biomedical domain-specific metric.
- **GPT-4o Score** (S_{GPT}): The trend continues with ProteinGPT_Mistral leading at 0.717, while ProteinGPT_LLaMA-3 scores 0.706, suggesting a close performance in this category.

In Figure 5, the performance of ProteinGPT with different base large language models (LLMs) is evaluated. The following findings are noteworthy:

- **ROUGE-1 and ROUGE-L**: The Mistral variant of ProteinGPT outperforms other base models (Vicuna, LLaMA-

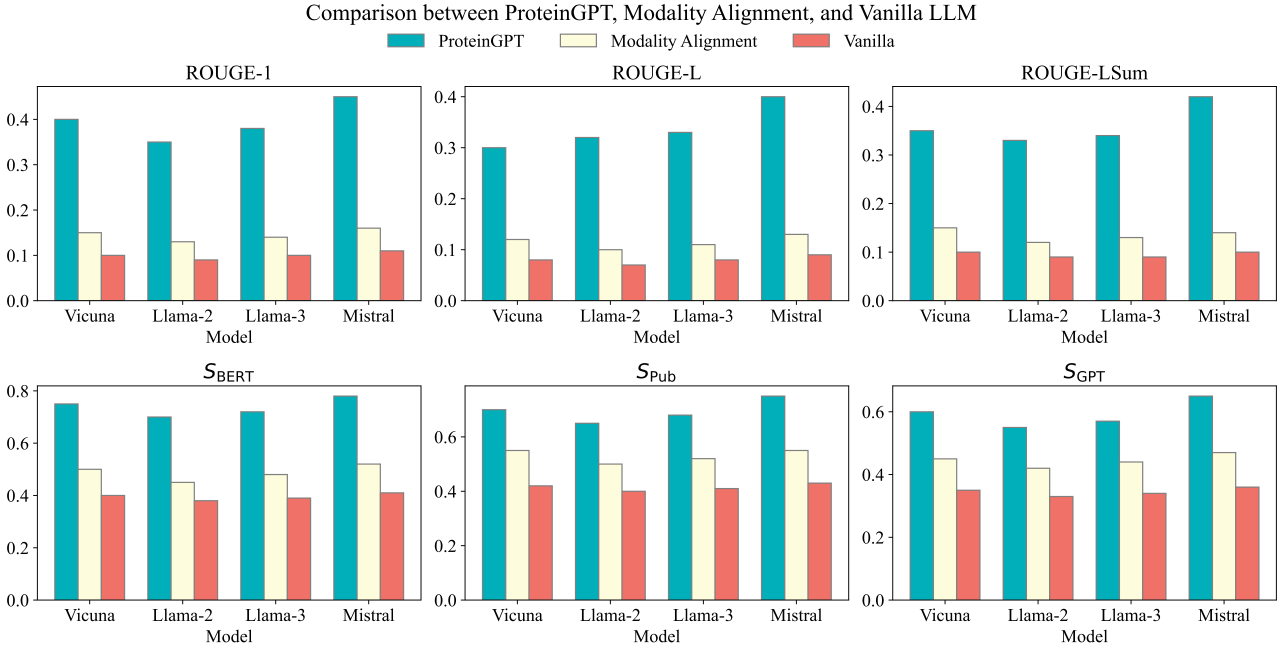
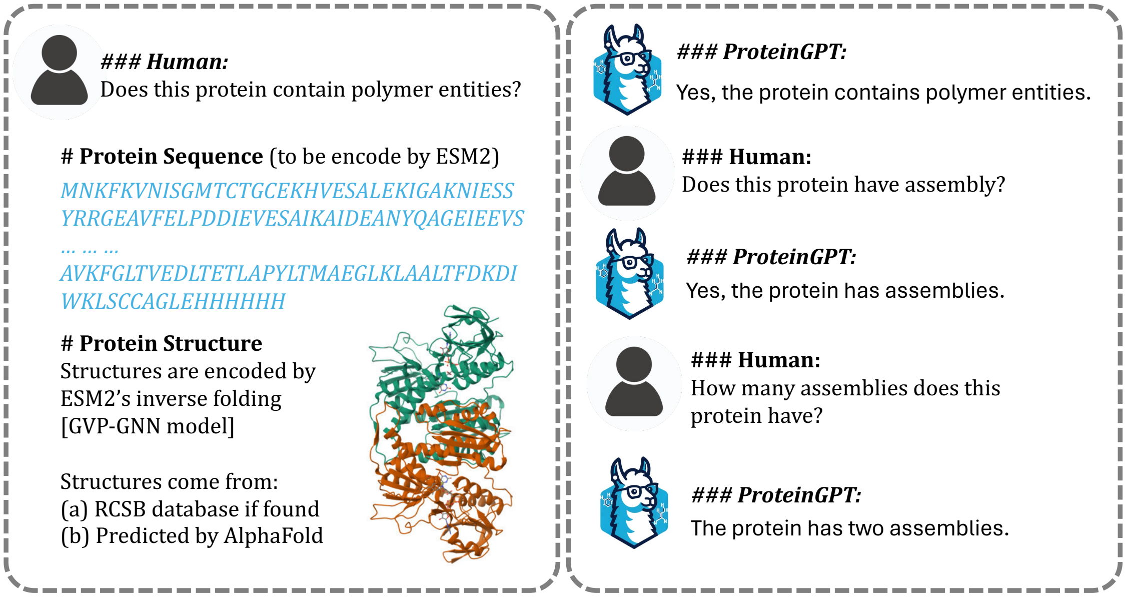


Figure 8: Comparison of Different Strategies and Models.

2, and LLaMA-3) in both ROUGE-1 (0.461) and ROUGE-L (0.460) scores. LLaMA-3 follows, but with slightly lower performance, while Vicuna and LLaMA-2 have the lowest scores.

- **ROUGE-LSum**: The Mistral base again outperforms others with a score of 0.457, while the LLaMA-3 variant is close behind at 0.367.
- **BERT Score**: Mistral shows the best performance (0.821), with other bases following in descending order: LLaMA-3 (0.779), LLaMA-2 (0.730), and Vicuna (0.739).
- **PubMedBert Score**: Mistral continues to outperform other base models with a score of 0.758, while LLaMA-3 follows at 0.706.
- **GPT-4o Score**: The Mistral base model scores highest at 0.717, followed by LLaMA-3 at 0.698.

Overall, the findings suggest that the ProteinGPT_Mistral variant consistently outperforms other models across all metrics. The Mistral base model proves to be the most effective for ProteinGPT, indicating that the combination of Mistral with ProteinGPT architecture is particularly powerful. In contrast, other base models such as LLaMA-2 and Vicuna, and the GPT variants, show comparatively lower performance, especially in metrics like ROUGE-1, ROUGE-L, and the PubMedBert Score, which are crucial for natural language processing tasks in specialized domains such as biomedical text analysis.



### Human:
Does this protein contain polymer entities?

# Protein Sequence (to be encode by ESM2)
MNKFKVNISGMTCTGCEKHVESALEKIGAKNIESS
YRRGEAVFELPDDIEVESAIKAIDEANYQAGEIEEVS
... ...
AVKFGLTVEDLTETLAPYLTMAEGLKLAALTFDKDI
WKLSCCAGLEHHHHHH

# Protein Structure
Structures are encoded by
ESM2's inverse folding
[GVP-GNN model]

Structures come from:
(a) RCSB database if found
(b) Predicted by AlphaFold

### ProteinGPT:
Yes, the protein contains polymer entities.

### Human:
Does this protein have assembly?

### ProteinGPT:
Yes, the protein has assemblies.

### Human:
How many assemblies does this protein have?

### ProteinGPT:
The protein has two assemblies.

Figure 9: ProteinGPT conversation on Protein 5X1Y.

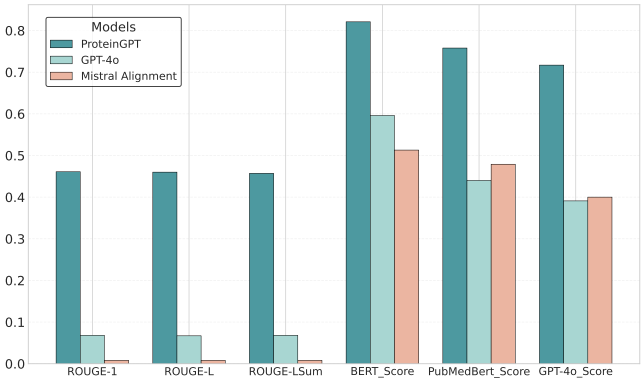
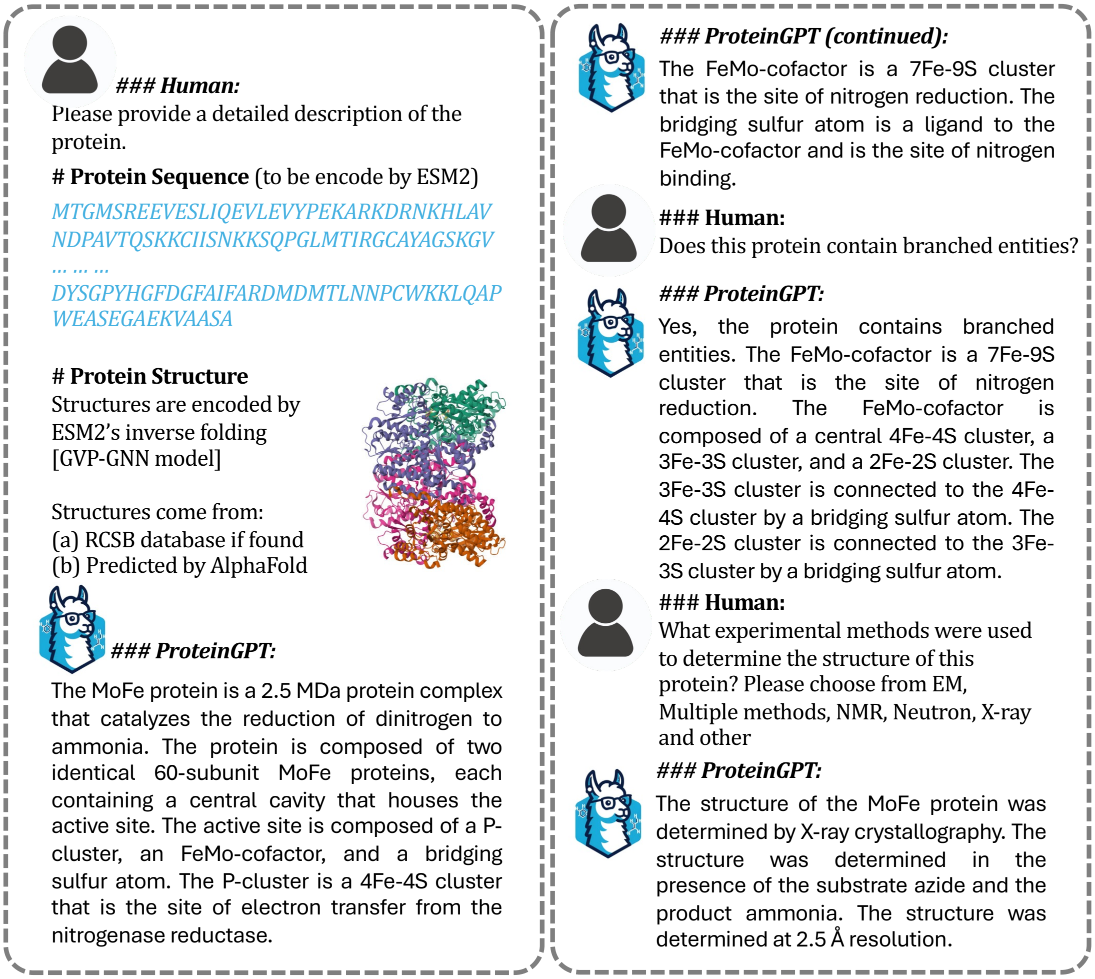


Figure 10: Comparison of ProteinGPT, OpenAI's GPT, and Modality-Aligned-only model.



### Human:

Please provide a detailed description of the protein.

# Protein Sequence (to be encode by ESM2)

MTGMSREEVESLIQEVLEVYPEKARKDRNKHLAV
NDPAVTQSKKCIISNKKSQPGLMTIRGCAYAGSKGV
… … …
DYSGPYHGFDGFAIFARDMDMTLNNPCWKKLQAP
WEASEGAEKVAASA

# Protein Structure

Structures are encoded by ESM2's inverse folding [GVP-GNN model]

Structures come from:
(a) RCSB database if found
(b) Predicted by AlphaFold

### ProteinGPT:

The MoFe protein is a 2.5 MDa protein complex that catalyzes the reduction of dinitrogen to ammonia. The protein is composed of two identical 60-subunit MoFe proteins, each containing a central cavity that houses the active site. The active site is composed of a P-cluster, an FeMo-cofactor, and a bridging sulfur atom. The P-cluster is a 4Fe-4S cluster that is the site of electron transfer from the nitrogenase reductase.

### ProteinGPT (continued):

The FeMo-cofactor is a 7Fe-9S cluster that is the site of nitrogen reduction. The bridging sulfur atom is a ligand to the FeMo-cofactor and is the site of nitrogen binding.

### Human:

Does this protein contain branched entities?

### ProteinGPT:

Yes, the protein contains branched entities. The FeMo-cofactor is a 7Fe-9S cluster that is the site of nitrogen reduction. The FeMo-cofactor is composed of a central 4Fe-4S cluster, a 3Fe-3S cluster, and a 2Fe-2S cluster. The 3Fe-3S cluster is connected to the 4Fe-4S cluster by a bridging sulfur atom. The 2Fe-2S cluster is connected to the 3Fe-3S cluster by a bridging sulfur atom.

### Human:

What experimental methods were used to determine the structure of this protein? Please choose from EM, Multiple methods, NMR, Neutron, X-ray and other

### ProteinGPT:

The structure of the MoFe protein was determined by X-ray crystallography. The structure was determined in the presence of the substrate azide and the product ammonia. The structure was determined at 2.5 Å resolution.

Figure 11: ProteinGPT's conversation on Protein 7RUV.

Metric	Base Model	(a) Protein Sequence S_{BERT}	(a) S_{Pub}	(a) S_{GPT}	(b) Modality Fusion S_{BERT}	(b) S_{Pub}	(b) S_{GPT}	(c) ProteinGPT S_{BERT}	(c) S_{Pub}	(c) S_{GPT}
Pre	Vicuna	0.572	0.464	0.396	0.582	0.515	0.446	0.730	0.626	0.605
	Llama-2	0.513	0.372	0.362	0.589	0.446	0.414	0.739	0.644	0.606
	Llama-3	0.490	0.442	0.369	0.593	0.487	0.446	0.779	0.706	0.698
	Mistral	0.525	0.405	0.362	0.513	0.479	0.400	0.821	0.758	0.717
Rec	Vicuna	0.653	0.473	0.310	0.691	0.540	0.334	0.788	0.751	0.661
	Llama-2	0.680	0.324	0.214	0.679	0.477	0.308	0.796	0.773	0.689
	Llama-3	0.657	0.332	0.210	0.695	0.456	0.309	0.803	0.776	0.713
	Mistral	0.624	0.287	0.192	0.623	0.359	0.211	0.839	0.816	0.752
F1	Vicuna	0.608	0.468	0.347	0.629	0.524	0.381	0.756	0.682	0.630
	Llama-2	0.582	0.345	0.269	0.628	0.459	0.351	0.764	0.701	0.644
	Llama-3	0.560	0.378	0.268	0.638	0.470	0.363	0.790	0.739	0.705
	Mistral	0.569	0.335	0.250	0.561	0.409	0.276	0.829	0.784	0.733

Table 3: Semantic Performance of LLM with Protein Sequence as Text Input (left), with Modality Fusion ONLY (middle), and with ProteinGPT (right). S_{BERT}, S_{Pub}, and S_{GPT} stand for BERTScore (Zhang et al., 2020), PubMedBERT-Score (Gu et al., 2021), and GPT-4o (OpenAI et al., 2024) score, respectively.

Model	**R-1**	**R-2**	**R-L**	**R-Lsum**
GPT-3.5-turbo	0.137	0.010	0.140	0.140
GPT-4-turbo	0.049	0.001	0.049	0.048
GPT-4o	0.068	0.000	0.067	0.068
OpenAI o1-mini	0.041	0.001	0.040	0.040
OpenAI o3-mini	0.072	0.007	0.072	0.073
Deepseek-R1	0.003	0.000	0.003	0.003
ProteinGPT$_{\text{Vicuna}}$	0.345	0.007	0.350	0.348
ProteinGPT$_{\text{Llama-2}}$	*0.348*	0.014	*0.347*	*0.349*
ProteinGPT$_{\text{Llama-3}}$	0.366	*0.021*	0.368	0.367
ProteinGPT$_{\text{Mistral}}$	**0.461**	**0.048**	**0.460**	**0.457**

Table 4: Lexical Performance with ProteinGPT, OpenAI's GPT/o-series models and DeepSeek R1 model. OpenAI and DeepSeek's models are text-based models. Therefore, protein sequences are fed into the model in text format, as part of the prompt.

Table 5: Lexical Performance of LLM with Protein Sequence as Text Input ONLY (Left), Modality Alignment ONLY (Middle), and ProteinGPT (Right). R-1, R-2, R-L, R-Lsum stand for ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, respectively.

	(a) Protein Sequence				(b) Modality Fusion				(c) ProteinGPT			
Model	R-1	R-2	R-L	R-Lsum	R-1	R-2	R-L	R-Lsum	R-1	R-2	R-L	R-Lsum
Vicuna	0.026	0.034	0.025	0.025	0.104	0.003	0.105	0.104	0.345	0.007	0.350	0.348
Llama-2	0.010	0.001	0.010	0.010	0.063	0.004	0.064	0.064	0.348	0.014	0.347	0.349
Llama-3	0.005	0.000	0.005	0.005	0.090	0.005	0.090	0.090	0.366	0.021	0.368	0.367
Mistral	0.007	0.001	0.007	0.007	0.008	0.002	0.008	0.008	0.461	0.048	0.460	0.457

G. Case Studies on OpenAI and DeepSeek models

G.1. OpenAI o1-mini Model

Below are OpenAI's o1-mini responses to questions on protein 6RE5's property and structure.

I. System Prompt

¡SYSTEM¿ You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. ¡/SYSTEM¿

II. Protein sequence information

`>6RE5_1|Chain A[auth 0]|ASA-10: Polytomella F-ATP synthase associated subunit 10|Polytomella sp. Pringsheim 198.80 (37502)`

MSYSAYFAKAGFQFPAGLSALVAGIVALNVCTGRPTKGTKEISNAEYNATPIGYLQSPDQHPTAFPKVPGMKDVHGSPHHHH

III. Question Group 1

```
> Question: How many assemblies does this protein have?
> Answer (o1-mini): The assembly state of the protein cannot be determined solely from
    the primary amino acid sequence provided.

> Question: How many entities does this protein have?
> Answer (o1-mini): The protein has **two entities**: the main protein sequence and the
     His-tag (HHHH).

> Question: How many heavy atom coordinates records does this protein have?
> Answer (o1-mini): The protein has **562** heavy atom coordinates records.

> Question: How many hydrogen atom coordinates records does this protein have?
> Answer (o1-mini): Hydrogen atom coordinates are not available from the primary amino
    acid sequence alone.

> Question: How many model structures deposited for this protein?
> Answer (o1-mini): As of October 2023, there are no model structures deposited for
    this protein sequence.

> Question: How many polymer monomers does this protein have?
> Answer (o1-mini): The protein consists of **82 amino acid monomers**.

> Question: How many heavy solvent atom coordinates records does this protein have?
> Answer (o1-mini): The number of heavy solvent atom coordinates records cannot be
    determined from the primary sequence alone.

> Question: How many nucleic acid polymer entities (DNA or RNA) does this protein have?
> Answer (o1-mini): The protein does not contain any nucleic acid polymer entities (DNA
     or RNA).

> Question: How many intermolecular covalent bonds does this protein have?
> Answer (o1-mini): This protein has **zero intermolecular covalent bonds**.

> Question: How many intermolecular metalic bonds does this protein have?
> Answer (o1-mini): Based on the provided sequence, the protein contains four histidine
     residues (`HHHH`) that are potential sites for intermolecular metal coordination
    bonds. Therefore, it can form up to **four intermolecular metal bonds**.

```

```
> Question: Does this protein contain non-polymer entities?
> Answer (o1-mini): No, based on the provided primary sequence, there are no indications of non-polymer entities.

> Question: Does this protein contain polymer entities?
> Answer (o1-mini): No, the protein does not contain polymer entities.

> Question: Does this protein contain DNA polymer entities?
> Answer (o1-mini): No, the provided protein sequence does not contain DNA polymerase domains.

> Question: Does this protein contain RNA polymer entities?
> Answer (o1-mini): No, this protein does not contain RNA polymerase domains.

> Question: Does this protein contain solvent entities?
> Answer (o1-mini): Yes, the protein contains residues (such as polar and charged amino acids) that are likely to interact with the solvent.

> Question: Does this protein contain branched entities?
> Answer (o1-mini): Yes, the protein contains several branched amino acids, including valine (V), isoleucine (I), and leucine (L).

> Question: Does this protein have unmodeled polymer monomers?
> Answer (o1-mini): Yes, the protein contains a His-tag sequence (HHHH) at the C-terminus, which may represent unmodeled polymer monomers.

> Question: Does this protein have hybrid nucleic acid polymer entities?
> Answer (o1-mini): No, based on the provided primary amino acid sequence, there is no indication of hybrid nucleic acid polymer entities.

> Question: Does this protein have cis-peptide linkages?
> Answer (o1-mini): Based solely on the primary sequence, the presence of proline residues (P) suggests potential sites for cis-peptide linkages. However, without structural data, it's not possible to definitively confirm the existence of cis-peptide bonds in this protein.
```

III. Question Group 2

```

> Question: What is the polymer entity composition for this protein?
> Answer (o1-mini): The protein consists of the following amino acids:

- **Alanine (A):** 10
- **Arginine (R):** 1
- **Asparagine (N):** 3
- **Aspartic acid (D):** 2
- **Cysteine (C):** 1
- **Glutamine (Q):** 3
- **Glutamic acid (E):** 2
- **Glycine (G):** 8
- **Histidine (H):** 6
- **Isoleucine (I):** 3
- **Leucine (L):** 4
- **Lysine (K):** 5
- **Methionine (M):** 2
- **Phenylalanine (F):** 4
- **Proline (P):** 8
- **Serine (S):** 6
- **Threonine (T):** 5
- **Tyrosine (Y):** 4
- **Valine (V):** 5

> Question: What is the nucleic acid polymer entity type for this protein?
> Answer (o1-mini): The nucleic acid polymer entity type for this protein is **messenger RNA (mRNA)**.

> Question: What are the bound nonpolymer components for this protein
> Answer (o1-mini): The protein likely binds nickel (\text{Ni}^{2+}) or cobalt (\text{Co}^{2+}) ions through its polyhistidine (HHHH) tag.

> Question: What experimental method(s) were used to determine the structure of this protein?
> Answer (o1-mini): Based solely on the provided primary amino acid sequence, the specific experimental method(s) used to determine the protein’s structure cannot be identified. Additional information, such as structural database entries or publication references, is required to ascertain the techniques employed.

> Question: What is the molecular mass (KDa) of polymer and non-polymer entities (exclusive of solvent) for this protein?
> Answer (o1-mini): The molecular mass of the protein polymer is approximately **8.82 kDa**. There are no significant non-polymer entities associated with this protein.

> Question: What is the polymer entity type for this protein?
> Answer (o1-mini): **Polymer Entity Type:** Polypeptide

> Question: What are the software programs reported in connection with the production of this protein?
> Answer (o1-mini): There are no specific software programs reported in connection with the production of this protein based solely on the provided primary sequence.

> Question: Is this protein determined by experimental or computational methods?
> Answer (o1-mini): It cannot be determined solely from the primary sequence.

> Question: What are the terms characterizing the protein?
> Answer (o1-mini): - **Polyhistidine Tag**: Contains a His\_4 tag at the C-terminus (HHHH).
- **Molecular Weight**: Approximately 10.7 kDa.
- **Isoelectric Point (pI)**: Around 6.2.
- **Amino Acid Composition**: Rich in alanine, glycine, and proline.
- **Secondary Structure**: Likely includes alpha helices and coiled regions.
- **Potential Functional Motifs**: May contain sites for phosphorylation or binding.
- **Solubility**: Predicted to be moderately soluble.
- **Expression Tag**: Suitable for affinity purification using nickel columns.

> Question: When is this protein first published?
> Answer (o1-mini): The first publication date of this protein cannot be determined solely from its primary amino acid sequence. Additional information such as the protein name, gene identifier, or related literature is required.
```

G.2. OpenAI o3-mini Model

Below are OpenAI’s o3-mini responses to questions on protein 1GC9’s property and structure.

I. System Prompt

¡SYSTEM¿ You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. ¡/SYSTEM¿

II. Protein sequence information

`>1GC9_1|Chain A|3-ISOPROPYLMALATE DEHYDROGENASE|Thermus thermophilus (300852)`

MKVAVLPGDGIGPEVTEAALKVLRALDEAEGLGLAYEVFPFGGAAIDAFGEPFPEPTRKGVEEAEAVLLGSVGG
PKWDGLPRKIRPETGLLSLRKSQDLFANLRPAKVFPGLERLSPLKEEIARGVDVLIVRELTGGIYFGEPRGMSEAE
AWNTERYSKPEVERVARVAFEGARKRRKHVVSVDKANVLEVGEFWRKTVEEVGRGYPDVALEHQYVDAMA
MHLVRSPARFDVVVTGNIFGDILSDLASVLPGSLGLLPSASLGRGTPVFEPVHGSAPDIAGKGIANPTAAILSAAM
MLEHAFGLVELARKVEDAVAKALLETPPPDLGGSAGTEAFTATVLRHLA

III. Question Group 1

```
> Question: How many assemblies does this protein have?
> Answer (o3-mini): It forms a single, monomeric assembly.

> Question: How many entities does this protein have?
> Answer (o3-mini): This protein is a single-chain entity.

> Question: How many heavy atom coordinates records does this protein have?
> Answer (o3-mini): There are 0 heavy atom coordinate records, as only the primary sequence is provided.

> Question: How many hydrogen atom coordinates records does this protein have?
> Answer (o3-mini): This protein’s primary sequence does not include any hydrogen atom coordinate records, so the answer is 0.

> Question: How many model structures deposited for this protein?
> Answer (o3-mini): There are no deposited model structures for this protein.

> Question: How many polymer monomers does this protein have?
> Answer (o3-mini): The protein consists of 349 amino acid residues.

> Question: How many heavy solvent atom coordinates records does this protein have?
> Answer (o3-mini): 0 heavy solvent atom coordinates records.

> Question: How many nucleic acid polymer entities (DNA or RNA) does this protein have?
> Answer (o3-mini): This protein contains no nucleic acid polymer entities.

> Question: How many intermolecular covalent bonds does this protein have?
> Answer (o3-mini): This protein has no intermolecular covalent bonds.

> Question: How many intermolecular metalic bonds does this protein have?
> Answer (o3-mini): 0

There is no evidence from the primary sequence indicating the formation of any intermolecular metallic bonds.

> Question: Does this protein contain non-polymer entities?
> Answer (o3-mini): No. The sequence only represents a polypeptide chain without any non-polymer entities.

> Question: Does this protein contain polymer entities?
> Answer (o3-mini): No, the protein is a single polypeptide chain composed solely of amino acids and does not include additional polymer entities.

> Question: Does this protein contain DNA polymer entities?
> Answer (o3-mini): No, this protein does not contain DNA polymer domains.

> Question: Does this protein contain RNA polymer entities?
> Answer (o3-mini): No, this protein does not seem to contain RNA polymerase domains.
```

III. Question Group 2

```

> Question: Does this protein contain solvent entities?
> Answer (o3-mini): No. The sequence lists only the amino acids without any solvent
    molecules or other non-protein entities.

> Question: Does this protein contain branched entities?
> Answer (o3-mini): No, this protein is a single linear polypeptide and does not
    contain branched entities.

> Question: Does this protein have unmodeled polymer monomers?
> Answer (o3-mini): No, the sequence appears complete with all expected polymer
    residues modeled.

> Question: Does this protein have hybrid nucleic acid polymer entities?
> Answer (o3-mini): No, this sequence represents a protein, not a hybrid nucleic acid
    polymer entity.

> Question: Does this protein have cis-peptide linkages?
> Answer (o3-mini): Based solely on the primary sequence, there’s no definitive
    evidence for cis-peptide bonds. Cis conformations - most often seen in X-Pro
    linkages-require structural data to confirm their presence.

> Question: What is the polymer entity composition for this protein?
> Answer (o3-mini): The protein is a single polypeptide chain composed of 346 standard
    L-amino acid residues.

> Question: What is the nucleic acid polymer entity type for this protein?
> Answer (o3-mini): This molecule is not a nucleic acid - it is a protein (an amino
    acid polymer).

> Question: What experimental method(s) were used to determine the structure of this
    protein?
> Answer (o3-mini): No experimental structure determination data is provided; only the
    primary sequence is given, so no method (e.g., X-ray crystallography or NMR
    spectroscopy) can be inferred.

> Question: What is the molecular mass (KDa) of polymer and non-polymer entities (
    exclusive of solvent) for this protein?
> Answer (o3-mini): The 344-residue chain weighs roughly 38 kDa, and there are no non-
    polymer entities present.

> Question: What is the polymer entity type for this protein?
> Answer (o3-mini): It is a protein, which is a polypeptide chain.

> Question: What are the software programs reported in connection with the production
    of this protein?
> Answer (o3-mini): No software programs were reported in connection with the
    production of this protein.

> Question: Is this protein determined by experimental or computational methods?
> Answer (o3-mini): There’s no accompanying annotation indicating experimental
    validation, so based solely on the primary sequence it’s unclear whether it was
```

```
determined experimentally or computationally.

> Question: When is this protein first published?
> Answer (o3-mini): The provided sequence does not include publication details, so the first publication date cannot be determined without additional reference data.
```

G.3. DeepSeek R1 Model

Below are DeepSeek's R1 model responses to questions on protein 5XSP's property and structure.

I. System Prompt

¡SYSTEM¿ You are an expert in protein domain, please answer the questions given the protein primary sequence and only give a concise answer to the question. ¡/SYSTEM¿

II. Protein sequence information

`>5XSP_1|Chains A, B|Phosphodiesterase acting on cyclic dinucleotides|Staphylococcus aureus (1280)`

GSMRTRVRARVISHALKDILAEGDKVIIMGHKRPDLDAIGAAIGVSRFAMMNNLEAYIVLNETDIDPTLRRVMN
EIDKKPELRERFITSDDAWDMMTSKTTVVIVDTHKPELVLDENVLNKANRKVVIDHHRRGESFISNPLLIYMEPY
ASSTAELVTELLEYQPTEQRLTRLESTVMYAGIIVDTRNFTLRTGSRTFDAASYLRAHGADTILTQHFLKDDVDT
YINRSELIRTVKVEDNGIAIAHGSDDKIYHPVTVAQAADELLSLEGIEASYVVARREDNLIGISARSLGSVNVQLT
MEALGGGGHLTNAATQLKGVTVEEAIAQLQQAITEQLSRSEDA

III. Question 1: Does this protein have cis-peptide linkages?

```
<think>
Okay, so I need to figure out whether the given protein has any cis-peptide linkages. I remember that in proteins, peptide bonds can be either cis or trans. The trans configuration is the most common because it’s more energetically favorable. Cis peptide bonds are less common but can occur under certain conditions, often when specific residues are involved.

Looking at the primary sequence provided, I should check for any residues that are known to favor cis peptide bonds. The key residues I remember are proline (Pro, P) and glycine (Gly, G). Proline is particularly known for its ability to adopt the cis configuration because of its ring structure, which restricts its conformational flexibility. Glycine, being the smallest amino acid with a lot of flexibility, can also sometimes be found in cis peptide bonds, though less commonly than proline.

I’ll scan through the sequence to spot any Pro or Gly residues. Let me go through each part:

- The sequence starts with:

GSMRTRVRARVISHALKDILAEGDKVIIMGHKRPDLDAIGAAIGVSRFAMMNNLEAYIVLNETDIDPTLRRVMNEIDKKPELRERFITSDDAWDMMTSKTT

DENVLNKANRKVVIDHHRRGESFISNPLLIYMEPYASSTAELVTELLEYQPTEQRLTRLESTVMYAGIIVDTRNFTLRTGSRTFDAASYLRAHGADTILTQ

SELIRTVKVEDNGIAIAHGSDDKIYHPVTVAQAADELLSLEGIEASYVVARREDNLIGISARSLGSVNVQLTMEALGGGGHLTNAATQLKGVTVEEAIAQL


Looking for Pro (P):

- I see a Pro at position ... let me count. Starting from the beginning: G, S, M, R, T, R, V, R, A, R, V, I, S, H, A, L, K, D, I, L, A, E, G, D, K, V, I, I, M, G, H, K, R
```

```
, P, D, L, D, A, I, G, A, A, I, G, V, S, R, F, A, M, M, N, N, L, E, A, Y, I, V, L, N, E, T, D, I, D, P, T, L, R, R, V, M, N, E, I, D, K, K, P, E, L, R, E, R, F, I, T, S, D, D, A, W, D, M, M, T, S, K, T, T, V, V, I, V, D, T, H, K, P, E, L, V, L, D, E, N, V, L, N, K, A, N, R, K, V, V, I, D, H, H, R, R, G, E, S, F, I, S, N, P, L, L, I, Y, M, E, P, Y, A, S, S, T, A, E, L, V, T, E, L, L, E, Y, Q, P, T, E, Q, R, L, T, R, L, E, S, T, V, M, Y, A, G, I, I, V, D, T, R, N, F, T, L, R, T, G, S, R, T, F, D, A, A, S, Y, L, R, A, H, G, A, D, T, I, L, T, Q, H, F, L, K, D, D, V, D, T, Y, I, N, R, S, E, L, I, R, T, V, K, V, E, D, N, G, I, A, I, A, H, G, S, D, D, K, I, Y, H, P, V, T, V, A, Q, A, A, D, E, L, L, S, L, E, G, I, E, A, S, Y, V, V, A, R, R, E, D, N, L, I, G, I, S, A, R, S, L, G, S, V, N, V, Q, L, T, M, E, A, L, G, G, G, G, H, L, T, N, A, A, T, Q, L, K, G, V, T, V, E, E, A, I, A, Q, L, Q, Q, A, I, T, E, Q, L, S, R, S, E, D, A.

So, I see several Pro residues. Let me note their positions:

1. The first Pro is at position 34 (counting from 0 or 1? Let me check: the sequence is 0-based or 1-based. Well, in the sequence given, the 34th character is P. Let's assume it's the 35th residue if starting count from 1.

Wait, actually, in the sequence provided, the 34th character is P. Let me count from the start:

G(1), S(2), M(3), R(4), T(5), R(6), V(7), R(8), A(9), R(10), V(11), I(12), S(13), H(14), A(15), L(16), K(17), D(18), I(19), L(20), A(21), E(22), G(23), D(24), K(25), V(26), I(27), I(28), M(29), G(30), H(31), K(32), R(33), P(34), D(35), L(36), D(37), A(38), I(39), G(40),
A(41), A(42), I(43), G(44), V(45), S(46), R(47), F(48), A(49), M(50), M(51), N(52), N(53), L(54), E(55), A(56), Y(57), I(58), V(59), L(60), N(61), E(62), T(63), D(64), I(65), D(66), P(67), T(68), L(69), R(70), R(71), V(72), M(73), N(74), E(75), I(76), D(77), K(78), K(79), P(80),
E(81), L(82), R(83), E(84), R(85), F(86), I(87), T(88), S(89), D(90), D(91), A(92), W(93), D(94), M(95), M(96), T(97), S(98), K(99), T(100), T(101), V(102), V(103), I(104), V(105), D(106), T(107), H(108), K(109), P(110), E(111), L(112), V(113), L(114), D(115), E(116), N(117), V(118), L(119), N(120),
K(121), A(122), N(123), R(124), K(125), V(126), V(127), I(128), D(129), H(130), H(131), R(132), R(133), G(134), E(135), S(136), F(137), I(138), S(139), N(140), P(141), L(142), L(143), I(144), Y(145), M(146), E(147), P(148), Y(149), A(150), S(151), S(152), T(153), A(154), E(155), L(156), V(157), T(158), E(159), L(160),
L(161), E(162), Y(163), Q(164), P(165), T(166), E(167), Q(168), R(169), L(170), T(171), R(172), L(173), E(174), S(175), T(176), V(177), M(178), Y(179), A(180), G(181), I(182), I(183), V(184), D(185), T(186), R(187), N(188), F(189), T(190), L(191), R(192), T(193), G(194), S(195), R(196), T(197), F(198), D(199), A(200),
A(201), S(202), Y(203), L(204), R(205), A(206), H(207), G(208), A(209), D(210), T(211), I(212), L(213), T(214), Q(215), H(216), F(217), L(218), K(219), D(220), D(221), V(222), D(223), T(224), Y(225), I(226), N(227), R(228), S(229), E(230), L(231), I(232), R(233), T(234), V(235), K(236), V(237), E(238), D(239), N(240),
G(241), I(242), A(243), I(244), A(245), H(246), G(247), S(248), D(249), D(250), K(251), I(252), Y(253), H(254), P(255), V(256), T(257), V(258), A(259), Q(260), A(261), A(262), D(263), E(264), L(265), L(266), S(267), L(268), E(269), G(270), I(271), E(272), A(273), S(274), Y(275), V(276), V(277), A(278), R(279), R(280),
E(281), D(282), N(283), L(284), I(285), G(286), I(287), S(288), A(289), R(290), S(291), L(292), G(293), S(294), V(295), N(296), V(297), Q(298), L(299), T(300), M(301), E(302), A(303), L(304), G(305), G(306), G(307), G(308), H(309), L(310), T(311), N(312), A(313), A(314), T(315), Q(316), L(317), K(318), G(319), V(320),
T(321), V(322), E(323), E(324), A(325), I(326), A(327), Q(328), L(329), Q(330), Q(331), A(332), I(333), T(334), E(335), Q(336), L(337), S(338), R(339), S(340), E(341), D(342), A(343).

So, Pro residues are at positions 34, 66, 79, 110, 141, 148, 165, 255.
```

```

Now, for each Pro, I should check the preceding residue because cis peptide bonds are
    between the carbonyl carbon of one residue and the amino nitrogen of the next. So,
    the Pro would be the residue after the peptide bond in question.

Wait, actually, the peptide bond is between the carboxyl of residue n and the amino of
    residue n+1. So, if residue n+1 is Pro, then the peptide bond between n and n+1
    could be cis.

So, I need to look for Pro at position n+1, and then check if the preceding residue (n)
     is such that a cis bond is favored.

But wait, the presence of Pro at position n+1 doesn't necessarily mean the bond is cis.
     It's just that Pro has a higher tendency to be in cis conformation when the
    preceding residue allows it. But in most cases, even with Pro, the bond is trans
    unless there's a specific reason for it to be cis, like in tight turns or specific
    structural motifs.

However, without structural data, it's hard to say definitively, but the presence of
    Pro residues increases the likelihood.

So, in the given sequence, there are multiple Pro residues. For each Pro at position n
    +1, the preceding residue is:

- Pro at 34: the preceding residue is R (Arginine) at 33.

- Pro at 66: preceding residue is D (Aspartic acid) at 65.

- Pro at 79: preceding residue is K (Lysine) at 78.

- Pro at 110: preceding residue is K (Lysine) at 109.

- Pro at 141: preceding residue is N (Asparagine) at 140.

- Pro at 148: preceding residue is Y (Tyrosine) at 147.

- Pro at 165: preceding residue is Q (Glutamine) at 164.

- Pro at 255: preceding residue is Y (Tyrosine) at 254.

Now, I need to consider if any of these preceding residues are Gly or Pro, but in this
    case, none of them are. However, Pro can sometimes form cis bonds even when the
    preceding residue is not Gly or Pro, especially in certain contexts like turns or
    loops.

But without the 3D structure, it's impossible to know for sure. However, the presence
    of multiple Pro residues suggests that there might be cis peptide bonds, but it's
    not a certainty. In the absence of specific structural information, we can't
    definitively say whether any of these Pro residues are in cis configuration.

Therefore, the answer would be that the protein may have cis-peptide linkages,
    especially around Pro residues, but without structural data, it's not certain.
</think>

The protein may have cis-peptide linkages, particularly around proline residues, but
    without structural data, this cannot be definitively confirmed.
```

III. Question 2: How many hydrogen atom coordinates records does this protein have?

```
<think>
Alright, let me try to figure this out. The question is asking how many hydrogen atom
    coordinates records a given protein has. The protein's primary sequence is provided
    , which is a long string of amino acids.

First, I know that each amino acid in a protein is made up of several atoms. Typically,
     an amino acid has a certain number of hydrogen atoms. For example, glycine, the
    simplest amino acid, has two hydrogen atoms in its side chain, while others might
    have more depending on their structure.

But wait, when we talk about the coordinates of hydrogen atoms in a protein structure,
    it's a bit different. In X-ray crystallography, which is commonly used to determine
     protein structures, hydrogen atoms aren't usually visible because they don't
    scatter X-rays strongly. So, their positions are often not recorded in the PDB (
    Protein Data Bank) files unless under certain conditions, like high-resolution
    structures or if they're part of a specific interaction.

Looking at the sequence provided, it's a long sequence, but without knowing the
    specific structure or the resolution of the data, it's hard to say exactly how many
     hydrogen coordinates there are. However, generally, most PDB entries don't include
     hydrogen atoms because they're not resolved. If they are included, it's usually
    because the structure was determined at a very high resolution, and the hydrogens
    were either located or modeled in.

So, considering standard practice, the answer is likely that the protein doesn't have
    any hydrogen atom coordinates recorded because they're not typically included in
    the PDB entries. Unless specified otherwise, most structures don't have hydrogen
    positions listed.
</think>

The protein does not have any hydrogen atom coordinates recorded. Hydrogen atoms are
    typically not included in standard PDB files due to their low visibility in X-ray
    crystallography.

Answer: 0
```

H. Related Works

Protein Representation Learning. Studies in protein representation mainly focus on the four hierarchical structural levels of proteins, which are critical for their biological roles. Some research (Madani et al., 2023; Notin et al., 2022) treat protein sequences as a biological language and utilize the Transformer architecture (Vaswani et al., 2017) to model amino acid interactions and predict sequences using large sequence databases. Other approaches (Lin et al., 2023; Meier et al., 2021; Rives et al., 2021; Rao et al., 2020; Vig et al., 2020) employ Masked Language Modeling (MLM) to develop attention mechanisms reflecting protein spatial interaction maps. Structure-oriented methods (Gligorijevi'c et al., 2021; Jing et al., 2020; Zhang et al., 2023c) encapsulate the functional attributes and spatial data of proteins for tasks like molecule binding (Jin et al., 2021; Kong et al., 2022), protein interface studies (Mahbub & Bayzid, 2022; R'eau et al., 2023), and property predictions (Zhang et al., 2022). However, most works rely on single-modal data, which overlooks the cross-modality interactions among text and protein sequence & structure information.

Large Language Models. Recent advancements in Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), Mistral Large 2 (Jiang et al., 2023), and Gemini (Team et al., 2023) have established new benchmarks in natural language processing (NLP), offering enhanced language comprehension and reasoning (Zellers et al., 2019; Wang et al., 2019; Frohberg & Binder, 2022; Yao et al., 2024; Zhang et al., 2023b). Multimodal LLMs (MLLMs) have further extended these capabilities beyond text, enabling the processing of natural language task performance on multimodal data (Shu et al., 2023; Yu et al., 2023; Ning et al., 2023). As proteins can be naturally represented by character strings, LLMs like ProteinChat (Guo et al., 2023) and ProtChatGPT (Wang et al., 2024) have been developed to effectively analyze protein structures and sequences.