# Generalize Large Language Model to Protein Understanding with GraphRAG

**Jingjie Zhang**[1], **Yiyi Zhang**[1], **Zijun Gao**[1], **Shaoning Li**[1]
[1] Department of Computer Science and Engineering, The Chinese University of Hongkong.
{1155224008, 1155229163, 1155224009, 1155229138}@link.cuhk.edu.hk

## Abstract

We propose **ProteinGraphRAG**, a graph-based retrieval-augmented generation (GraphRAG) approach tailored to protein-centric tasks. By leveraging GraphRAG, ProteinGraphRAG does not require extensive fine-tuning on complex protein contexts. Instead, it aids general large language models (LLMs) in comprehensively understanding protein-related entities and their interconnections, thereby enabling more accurate and informed decision-making.

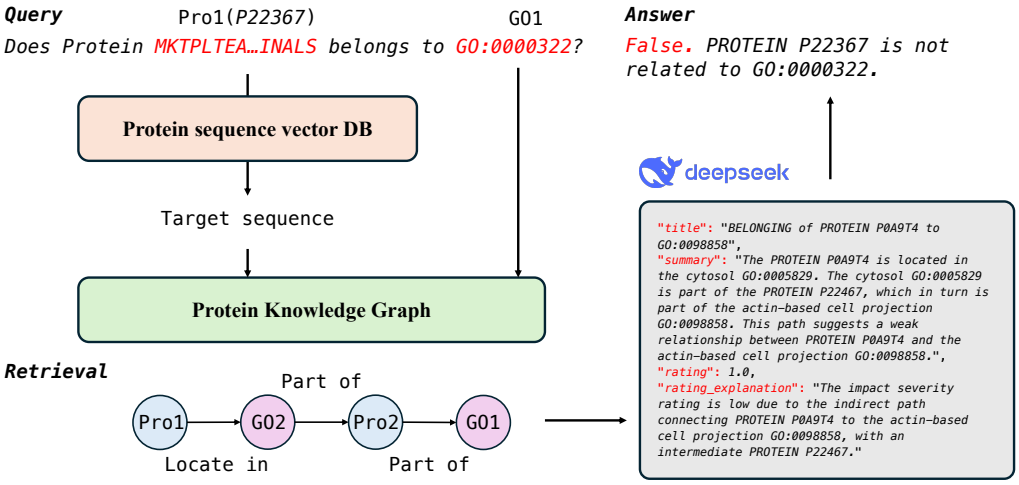


Figure 1: **Schematic overview of the ProtGraphRAG pipeline.** A query specifying a source protein and a target Gene Ontology (GO) term is first decomposed to identify the relevant protein entity and GO identifier. These entities and their interconnections are then retrieved from the constructed protein knowledge graph. Next, the retrieved information is provided to a large language model (e.g., deepseek) to generate a context-aware summary. Finally, the LLM utilizes this summary to determine whether the source protein belongs to the target GO term.

## 1 Introduction

In this technical report, we present a method for enabling general large language models (LLMs) to perform protein-centric tasks such as Gene Ontology (GO) classification via GraphRAG, without

Preprint. Under review.

specialized fine-tuning on protein sequences. The overall pipeline is illustrated in Figure 1. First, we construct a protein knowledge graph derived from the large-scale ProteinKG25 [2], which encompasses 590,914 entities and 5,099,141 different relations between proteins and GO terms. We then leverage ESM C [1] to embed protein sequences, producing 960-dimensional vectors. Once the sequence most similar to a query protein is identified, we employ a shortest-path algorithm to extract the path linking the source protein and the target GO term. The relevant entities and relations along this path are then passed to an LLM to generate a structured summary, which includes a title, a concise synopsis, a "belonging" rating, and an explanation for this rating. Finally, the LLM uses this summary to make a definitive decision regarding the source protein's membership in the specified GO term. The project is based on nano-graphrag[1].

The contributions are summarized as follows:

- First application of GraphRAG to protein-centric tasks. We introduce a novel approach that leverages GraphRAG on the large-scale protein knowledge graph to achieve protein classification and related tasks.
- Elimination of specialized protein language models or custom tokens. Our method obviates the need for pre-trained protein language models or adding specialized tokens, allowing easy integration with diverse LLM/APIs without additional fine-tuning.
- Improved performance over existing protein language models. Using our ProteinGraghRAG, the proposed method outperforms conventional protein language models in key protein-centric tasks.

## 2 Experiments

To evaluate the robustness of our proposed ProteinGraphRAG method against conventional protein language models, we selected ten proteins not included in the knowledge graph. For each of these ten proteins, we compared with ProtLLM [3] approach: a protein language model pretrained on UniRef50 and fine-tune on GO tasks. The resulting metrics: Accuracy, F1-score, AUC, Recall, and Precision are displayed in Fig. 2.

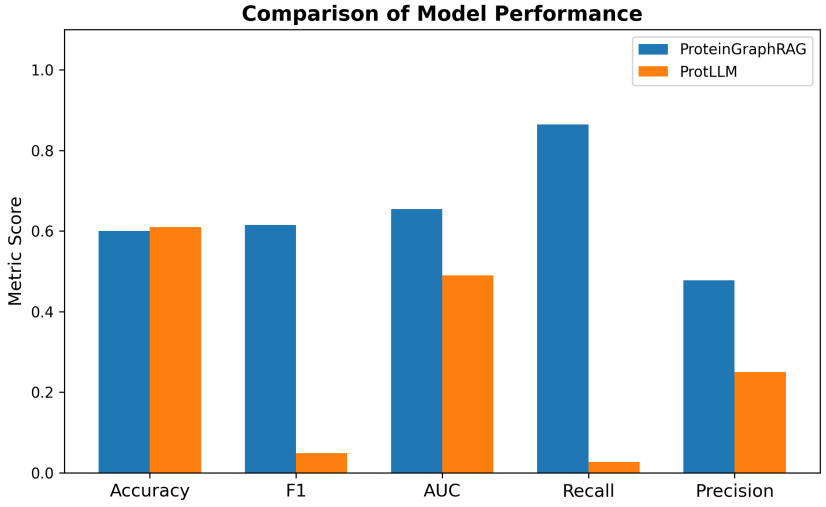


Figure 2: Comparison of performance metrics for ProteinGraphRAG and ProtLLM.

Although the two methods showed similar accuracy, ProteinGraphRAG had substantially higher F1, AUC, recall, and precision. This indicates that graph-based retrieval adds valuable relational context, leading to more reliable protein–GO predictions—particularly by reducing missed associations and improving overall confidence in the results.

[1] https://github.com/gusye1234/nano-graphrag

## References


[1] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

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

[3] L. Zhuo, Z. Chi, M. Xu, H. Huang, H. Zheng, C. He, X.-L. Mao, and W. Zhang. Protllm: An interleaved protein-language llm with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920*, 2024.