Proteomics Chatbot

Dipesh Tripathi
Computer Science Department
University of South Dakota
Vermillion, SD, USA
dipesh.tripathi@coyotes.usd.edu

Sumaly Bajracharya
Computer Science Department
University of South Dakota
Vermillion, SD, USA
sumaly.bajracharya@coyotes.usd.edu

Ujjwal Bhatta
Computer Science Department
University of South Dakota
Vermillion, SD, USA
ujjwal.bhatta@coyotes.usd.edu

Jeevan Kaphle
Computer Science Department
University of South Dakota
Vermillion, SD, USA
jeevan.kaphle@coyotes.usd.edu

***Abstract*—Extracting relevant protein information in a natural language is a challenging task. This report presents the information retrieval for protein from the UniProt database using our Proteomics Chatbot interface. Four distinct approaches are implemented and evaluated: (1) embedding protein records using a SentenceTransformer, (2) ProtBERT for protein-specific embeddings, (3) directly retrieving protein data via the UniProt API by extracting the primary accession (UniProt ID), and (4) a hybrid approach - combining FAISS-indexed vector search with API retrieval using the Llama3.2:1b language model. The first three approaches yield valuable insights and highlight specific challenges, which directs us to the design of our final hybrid solution. Approach 4 successfully balances flexibility and accuracy, achieving high accuracy across diverse query types. The hybrid approach correctly identifies and retrieves information about specific proteins even when queries are formulated with descriptive terms rather than explicit IDs.**

***Index Terms*—SentenceTransformer, ProtBERT transformer, direct API, hybrid, Retrieval-Augmented Generation (RAG), FAISS index.**

I. Introduction

The Proteomics Chatbot uses a conversational, natural language interface to get relevant protein information from the UniProt database. By enabling users to retrieve detailed protein data with simple queries, the project aims to bridge the gap between complex bioinformatics databases and user-friendly applications.

As data continues to increase in volume and complexity, resources like UniProt offer an abundance of information through their portals. However, the template-driven nature of search interfaces that utilize identifiers, filters, and predefined domain knowledge can be difficult to navigate, even for experts. This presents a challenge for early-stage learners and researchers, interdisciplinary scientists, or domain experts outside the field who are not trained in formal query languages.

This project leverages large language models (Llama3.2:1b) combined with modern retrieval techniques and deep learning-based embeddings to address this challenge. We experiment with both general-purpose sentence embeddings and domain-specific models like ProtBERT to index protein records. In addition, direct API allows for exact matches for precision.

The main objective is to deliver accurate, contextually rich responses while supporting follow-up interactions in a conversational format. The Proteomics Chatbot aims to give researchers, bioinformaticians, and proteomics students an approachable tool by utilizing technologies like transformer-based models and FAISS indexing.

Some recent works, such as ExpasyGPT [1] and ProteinChat [2], which integrate large language models with structured biomedical data to enhance accessibility, accuracy, and usability in proteomics research, have been done. However, in our project, we have combined the FAISS-indexed vector with API retrieval to produce more accurate results than the rest of the approaches.

II. Literature Review

Retrieving accurate protein information from large biological databases like UniProt has become increasingly feasible with recent advancements in natural language processing (NLP) and information retrieval. Traditional keyword-based search techniques often fail when users provide incomplete or descriptive queries, requiring more flexible approaches.
Sentence-Transformers in [3] have emerged as a powerful method for generating dense semantic embeddings of text, allowing efficient similarity search. Combined with FAISS [4], these embeddings enable rapid retrieval of relevant protein records based on meaning rather than exact keyword matches. However, general-purpose embeddings may not fully capture protein-specific terminologies. To address this, domain-tuned models like ProtBERT, trained on billions of protein sequences, have shown superior ability to represent protein features like motifs, structures, and biological functions [5]. ProtBERT helps handle biological terms better, though multi-field protein records still present challenges for single-vector embeddings.
Recent literature [6] highlights the rise of Retrieval-Augmented Generation (RAG), where large language models (LLMs) such as LLaMA or GPT-4 are paired with retrieval systems for accurate fact-based responses. In biomedical domains, RAG has demonstrated significant success in reducing

hallucinations and improving reliability [7].

Moreover, symbolic retrieval approaches, such as direct API lookups using UniProt IDs, offer unmatched precision when specific identifiers are available [8]. However, this method lacks flexibility for vague or exploratory queries. As a result, hybrid retrieval systems combining semantic similarity search with exact identifier lookups are emerging as the most effective solution [9]. This design ensures flexible query handling with semantic embeddings while leveraging the accuracy of database-backed API results.

The integration of such hybrid approaches in proteomics chatbots ensures better coverage, accuracy, and user satisfaction, aligning well with trends in modern AI-driven biomedical information systems.

Moreover, there is awareness that most recent works focus on qualitative evaluations and relevance feedback rather than standard NLP evaluation metrics.

III. Data

A. Dataset Overview

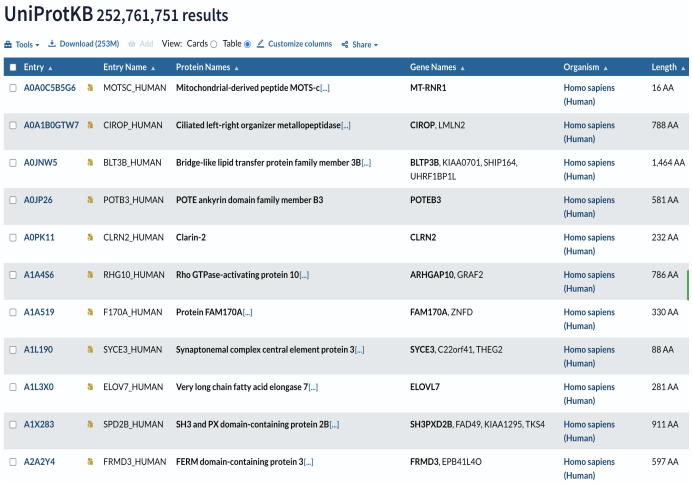
The primary dataset for the project is derived from the UniProt database [10]—a comprehensive resource for protein sequence and annotation data (around 253 million records). UniProt provides detailed records for millions of proteins, including unique identifiers (primaryAccession and UniProtKB ID), descriptive annotations, gene information, and functional details. Each protein entry in UniProt is structured to include multiple fields.

- **UniProt data**: https://www.uniprot.org/uniprotkb?query=.
- **Data taken**: 500.
- **Key fields**: 'primaryAccession', 'proteinName', 'geneName', 'organism', 'accessions', 'proteomeId', 'function'.
- **Field datatype**: string.

The key fields used to extract protein information in our project:

- **Primary accession**: A unique identifier for each protein record.
- **Protein name**: Consists of the description, recommended name, and full name of the protein.
- **Gene names**: Consists of the gene name of the protein, typically provided as an array of objects, concatenated with commas.
- **Organism name**: Consists of the organism name or the scientific name of the protein.
- **Accession**: Consists of the secondary accession of the protein.
- **Proteome**: Proteome IDs where the protein is found.
- **Functional annotations**: Functional description of the protein, extracted from the function comment type.

The UniProt data is initially stored as a CSV file (uniprot_extracted_data.csv). This file serves a dual purpose: first, it is utilized to build FAISS indices by extracting and concatenating relevant fields from each record; second, it supports direct API queries for retrieving up-to-date information as needed. By leveraging the structured nature of the UniProt records, the project efficiently combines offline vector-based retrieval with real-time API calls, ensuring both speed and accuracy in responding to user queries.



UniProtKB 252,761,751 results

Tools · Download (253M) · Add · View: Cards ○ Table ◉ · Customize columns · Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[...]	MT-RNR1	Homo sapiens (Human)	16 AA
A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metallopeptidase[...]	CIROP, LMLN2	Homo sapiens (Human)	788 AA
A0JNW5	BLT3B_HUMAN	Bridge-like lipid transfer protein family member 3B[...]	BLTP3B, KIAA0701, SHIP164, UHRF1BP1L	Homo sapiens (Human)	1,464 AA
A0JP26	POTB3_HUMAN	POTE ankyrin domain family member B3	POTEB3	Homo sapiens (Human)	581 AA
A0PK11	CLRN2_HUMAN	Clarin-2	CLRN2	Homo sapiens (Human)	232 AA
A1A4S6	RHG10_HUMAN	Rho GTPase-activating protein 10[...]	ARHGAP10, GRAF2	Homo sapiens (Human)	786 AA
A1A519	F170A_HUMAN	Protein FAM170A[...]	FAM170A, ZNFD	Homo sapiens (Human)	330 AA
A1L190	SYCE3_HUMAN	Synaptonemal complex central element protein 3[...]	SYCE3, C22orf41, THEG2	Homo sapiens (Human)	88 AA
A1L3X0	ELOV7_HUMAN	Very long chain fatty acid elongase 7[...]	ELOVL7	Homo sapiens (Human)	281 AA
A1X283	SPD2B_HUMAN	SH3 and PX domain-containing protein 2B[...]	SH3PXD2B, FAD49, KIAA1295, TKS4	Homo sapiens (Human)	911 AA
A2A2Y4	FRMD3_HUMAN	FERM domain-containing protein 3[...]	FRMD3, EPB41L4O	Homo sapiens (Human)	597 AA

Fig. 1. UniProt data

Figure 1 shows UniProt data that has been used to retrieve information about the protein.

IV. Methodology

The methodology for this project involves using four approaches: sentence transformers for embedding entire rows, protbert transformer for protein-specific embedding, direct API calls using UniProt ID, and a hybrid approach - combining vector search with API retrieval.

1) **Approach 1: Sentence transformer for embedding entire rows:**
   The SentenceTransformer method applies a universal text embedding model featuring a protein record transformer to embed proteins into high-dimensional vectors. In this approach, the all-MiniLM-L6-v2 model, which is known for producing 384 embeddings that consist of text segments, is used. Each protein entry contains multiple fields (primary accession, protein name, gene name, organism name, accessions, proteome, and function), which are transformed into a single string and embedded as a vector. The stored vectors are indexed using FAISS, a library developed by Facebook AI, which helps search and cluster dense vectors.
   This method's benefit is its effectiveness in dealing with natural language questions without needing precise protein ID. Having both the query and protein records stored in vector space enables the discovery of proteins that are still relevant to the query, even if searched partially. RAG (Retrieval-Augmented Generation) architecture provides accurate and contextual responses to user queries.
   RAG architecture: The system applies the RAG model, where information retrieval is done alongside text generation for precise and contextual responses. RAG Components:

- **Retriever:**
  - Uses FAISS (Facebook AI Similarity Search) for efficient vector similarity search.
  - Embeddings are generated using Sentence Transformers (all-MiniLM-L6-v2).
  - Enables semantic search beyond simple keyword matching.
- **Generator:**
  - Utilizes the Llama model for natural language generation.
  - Augments responses with the retrieved context.

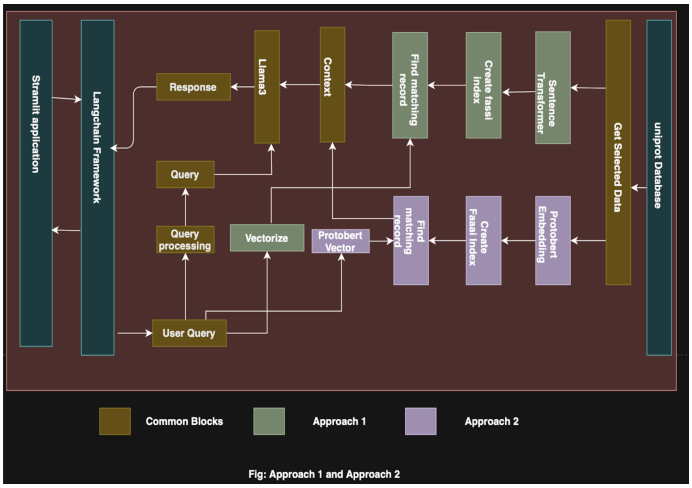


Fig. 2. Approach 1 and 2

2) **Approach 2: ProtBERT transformer for protein-specific embedding:**
The ProtBERT model takes a domain-specific approach to a transformer model pre-trained on over 6 billion protein sequences. ProtBERT was designed to understand the language of proteins, including their unique vocabulary, structural patterns, and evolutionary relationships. This model produces embeddings that are relatively more aligned to the protein's attributes and so, is expected to capture the detailed interrelationships among protein entries better.
Similar to Approach 1, we embed entire protein records by concatenating multiple fields into a single text representation. However, ProtBERT processes this information specifically for protein data. These embeddings are also indexed with FAISS in order to perform fast similarity searches on them.
Figure 2 gives the architecture for the sentence transformer and the protbert transformer embedding.
3) **Approach 3: Direct API calls using UniProt ID:** The third approach focuses on ID extraction from the query, which is more precise on the retrieval rather than semantic similarity. It applies a model Llama3.2:1b to the user's queries and analyzes them to extract possible UniProt identifiers (primaryAccession). After a valid ID is recognized, the system completely avoids vector similarity search, and instead it uses a direct API call to the UniProt database for the complete protein record. This exact record contains all pertinent current details and is valid.
This approach does not have flexibility because every time the user wants to get a certain protein by its ID, they should be able to query its ID and be guaranteed that they will obtain the right actual information without being misled by other similar proteins. Figure 3 presents the architecture for retrieving protein information using direct API calls.

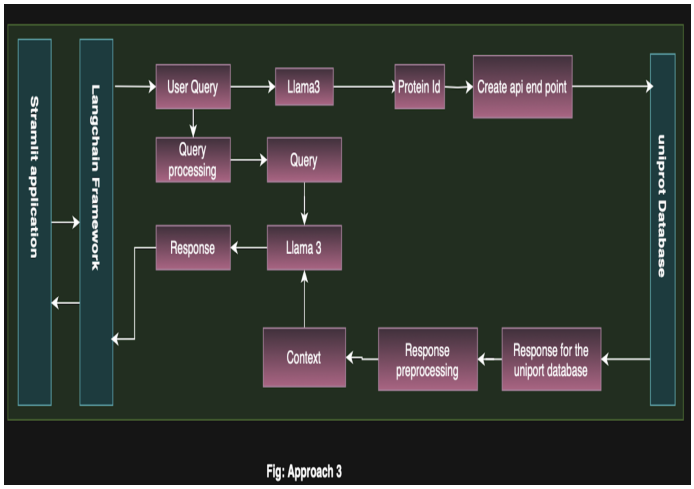


Fig. 3. Direct API calls using UniProt ID

4) **Approach 4: Hybrid approach - combining vector search with API retrieval:** The hybrid approach combines the strengths of vector-based similarity search with direct API retrieval to overcome the limitations of previous methods. This approach first leverages Sentence-Transformer embeddings to identify the most relevant protein record based on semantic similarity, then extracts the primaryAccession from this record to perform a precise API call to the UniProt database.
For each user query, the system first processes and vectorizes it using the same embedding model applied to the protein records. It then searches the FAISS index to find the closest matching protein entry. Rather than returning this potentially incomplete or outdated information directly, the system extracts the primaryAccession ID and uses it to query the UniProt API for the most current and comprehensive protein data.
The methodology of the hybrid approach is summarized in Figure 4.

V. Experiments and Results

A. Experimentation Overview

In this study, we conducted experiments to retrieve protein information using four approaches. The experiments were carried out on a Google Colab environment utilizing an NVIDIA T4 GPU.

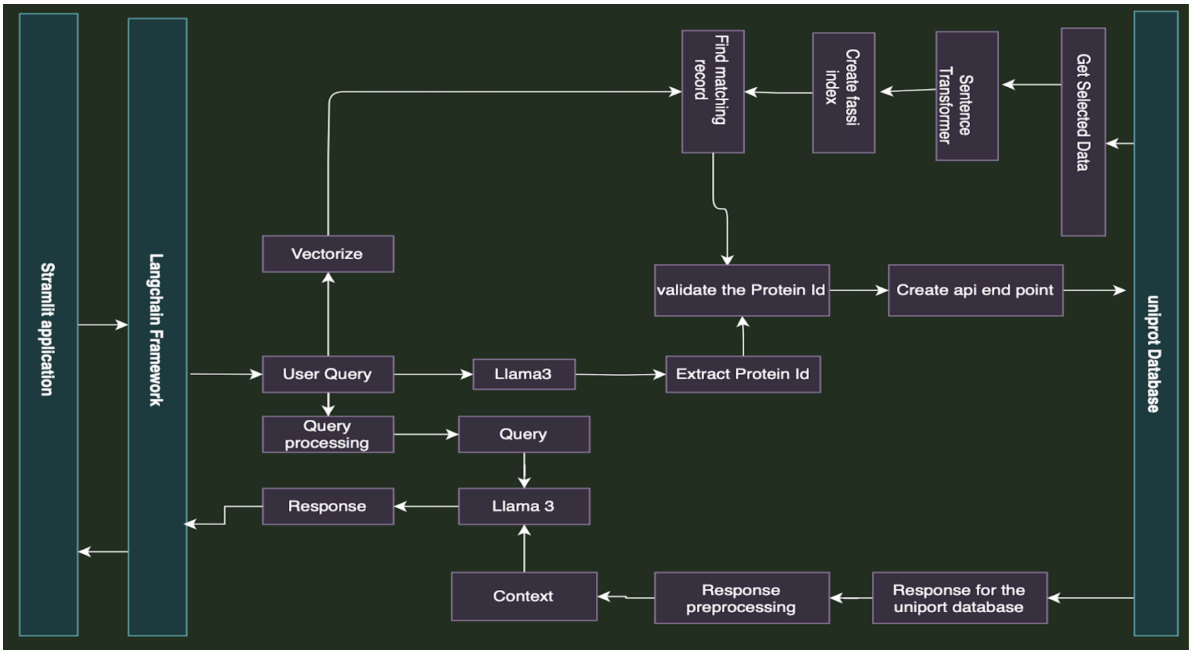


Fig. 4. Approach 4: hybrid approach

Embedding models: SentenceTransformer (all-MiniLM-L6-v2), ProtBERT transformer
Language model: Ollama (Llama3.2:1b)
Index: FAISS for fast similarity search
Architecture: Retrieval-Augmented Generation

B. Results

Approach 1: Sentence transformer
Experiment 1:
Query: *Give me the function of mitochondrial-derived peptide mots-c*
Context: *['id': 'O14925', 'protein_name': 'Mitochondrial import inner membrane translocase subunit Tim23', 'gene_names': 'TIMM23', 'organism_name': 'Homo sapiens', 'accessions': 'Q53FF8, Q5T1E6, Q6P5S5', 'proteome': 'UP000005640', 'function': 'Essential component of the TIM23 complex, a complex that mediates the translocation of transit peptide-containing proteins across the mitochondrial inner membrane']*
Response: *The protein Tim23 is involved in the import of inner membrane translocon peptides across the mitochondria. It specifically mediates the translocation of transit peptide-containing proteins into the mitochondrial matrix.*

Approach 2: ProtBERT transformer
Experiment 2:
Query: *Give me the function of mitochondrial-derived peptide mots-c*
Context: *['id': 'O14925', 'protein_name': 'Mitochondrial import inner membrane translocase subunit Tim23', 'gene_names': 'TIMM23', 'organism_name': 'Homo sapiens', 'accessions': 'Q53FF8, Q5T1E6, Q6P5S5', 'proteome': 'UP000005640', 'function': 'Essential component of the TIM23 complex, a complex that mediates the translocation of transit peptide-containing proteins across the mitochondrial inner membrane']*
Response: *The mitochondrial-derived peptide MOTS-c is involved in regulating insulin sensitivity and metabolic homeostasis through its inhibition of the folate cycle, thereby reducing de novo purine biosynthesis which leads to the accumulation of 5-aminoimidazole-4-carboxamide (AICAR) and activation of 5'-AMP-activated protein kinase (AMPK).*

Approach 3: Direct API calls using UniProt ID
Experiment 3:
Query: *Give me the function of the following protein: A0A0C5B5G6*
Formulated query to get the protein ID: *The provided sentence is "give me the function of following protein: A0A0C5B5G6". Extract only the protein id from the given sentence if present. If the given sentence is referring to the previous chat history, then provide the response as "TRUE". If the given sentence does not contain protein id then give "No protein id found" as response. Note: only give the protein ID, only one word for the protein ID, and do not give any other words in the response.*
Protein id response from the model: *"protein name is A0A0C5B5G6"*
Context: *'entryType': 'UniProtKB reviewed (Swiss-Prot)', 'primaryAccession': 'A0A0C5B5G6', 'secondaryAccessions': [],.....Reverses VC-induced reduction in AMPK phosphorylation and decreases expression of receptors AGTR1 and EDNRB (PubMed:31694019)']*
Response: *Based on the UniProtKB accession code*

A0A0C5B5G6, the protein in question is called the Mitochondrial-derived peptide MOTS-c. The protein MotS-c is involved in regulating mitochondrial function and metabolism..... metabolic stress, and promoting cellular survival and growth.

Approach 4: Hybrid approach

Experiment 4:

Queries and responses for the hybrid approach are shown in the figure 5 and 6 respectively.

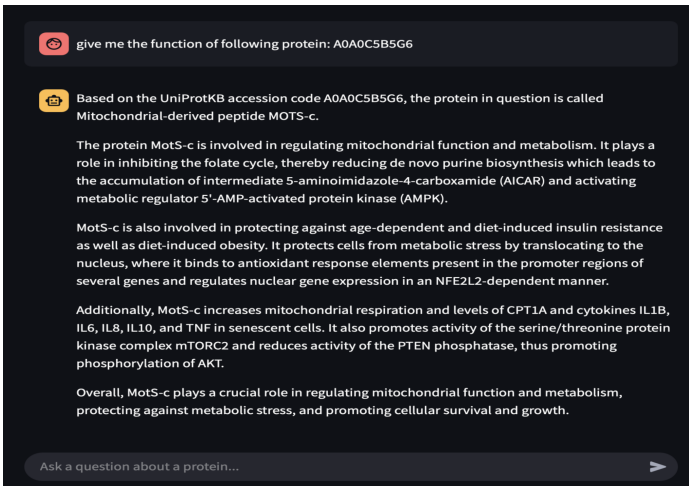


Fig. 5. Response from hybrid approach

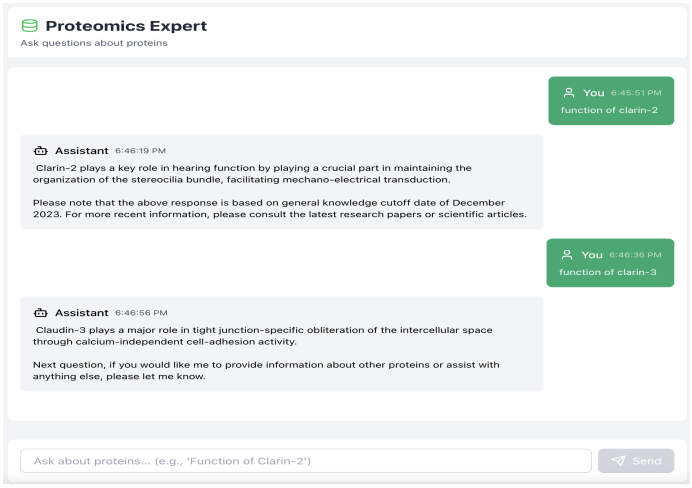


Fig. 6. Response from approach 4

For Figure 5:

Context: *['id': 'A0A0C5B5G6', 'protein_name': 'Mitochondrial-derived peptide MOTS-c', 'gene_names': 'MT-RNR1', 'organism_name': 'Homo sapiens', 'accessions': '', 'proteome': 'UP000005640', 'function': 'Regulates insulin sensitivity and metabolic homeostasis (PubMed:25738459, PubMed:33468709).....Promotes osteoblast proliferation and osteoblast synthesis of type I collagens COL1A1 and COL1A2 via the TGFB/SMAD pathway (PubMed:31081069)']*

For Figure 6:

Context: *[{'id': 'A0PK11', 'protein_name': 'Clarin-2', 'gene_names': 'CLRN2', 'organism_name': 'Homo sapiens', 'accessions': nan, 'proteome': 'UP000005640', 'function': 'Plays a key role to hearing function. Required for normal organization and maintenance of the stereocilia bundle and for mechano-electrical transduction', {'id': 'O15551', 'protein_name': 'Claudin-3', 'gene_names': 'CLDN3', 'organism_name': 'Homo sapiens', 'accessions': nan, 'proteome': 'UP000005640', 'function': 'Barrier-forming claudin. Plays a major role in tight junction-specific obliteration of the intercellular space, through calcium-independent cell-adhesion activity'....to contribute to the recognition of the similarity)']}*

VI. Discussion

The analysis of the results from the language models revealed several important insights. Protein information retrieval based on vector similarity is fast, but it sometimes fails to capture the subtle relationships among the diverse protein attributes. While the system successfully retrieves protein records, the system would sometimes return proteins that shared terminology but had entirely different functional roles. The term *"mitochondrial"* in our query matches with *mitochondrial import inner membrane translocase subunit Tim23* rather than *"mitochondrial-derived peptide"*, which generates false context, eventually resulting in an incorrect response. ProtBERT embeddings capture domain-specific features better, but challenges remain with multi-column relationship handling, which leads to inconsistent match accuracy. While ProtBERT excels at biological pattern recognition by making domain-appropriate inferences (correctly mentioning metabolic regulation for MOTS-c), it still fails to properly weight identifier fields, leading to incorrect protein retrieval. Approach 3 is highly accurate for queries that include a specific UniProt ID; however, it lacks flexibility when users enter queries without an explicit ID. Approach 4 achieves excellent precision across a variety of query types by effectively showing a balance between accuracy and flexibility. Even when queries are written using descriptive phrases rather than explicit IDs, the hybrid technique is still able to correctly identify and retrieve information about particular proteins. Since the context obtained for a particular query is accurate, as seen in figure 5 and 6, the response obtained is also precise and relevant.

VII. Conclusion and Future Direction

A. Conclusion

This project highlighted the trade-offs between flexibility and accuracy in different approaches. While direct API calls ensure correct retrieval when a UniProt ID is provided, they limit query flexibility. On the other hand, embedding-based retrieval offers flexibility but struggles with consistently capturing all nuanced relationships among fields. The combination

of vector search and API retrieval appeared promising, as it leveraged the advantages of both strategies, giving the best result. Based on the promising results of our hybrid approach, our future work will focus on comprehensive evaluation, optimization, and expansion of the system. To support comprehensive evaluation, we plan to implement practical tools such as Top-k Retrieval Accuracy (Hit@k) for ranking performance.

REFERENCES

[1] SIB Swiss Institute of Bioinformatics, "Expasygpt: Ai-driven query tool for life science databases," https://www.expasy.org/chat, 2025, accessed: 2025-05-01.

[2] M. Huo, H. Guo, X. Cheng, D. Singh, H. Rahmani, S. Li, P. Gerlof, T. Ideker, D. A. Grotjahn, E. Villa, L. Song, and P. Xie, "Multi-modal large language model enables protein function prediction," *bioRxiv*, 2024.

[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 3982–3992.

[4] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *arXiv preprint arXiv:1702.08734*, 2017.

[5] A. Elnaggar, M. Heinzinger, C. Dallago *et al.*, "Prottrans: Towards understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7112–7127, 2021.

[6] P. Lewis, E. Perez, A. Piktus *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[7] M. Li, H. Kilicoglu, H. Xu, and R. Zhang, "Biomedrag: A retrieval-augmented large language model for biomedicine," *arXiv preprint arXiv:2405.00465*, 2024.

[8] Q. Jin, Y. Yang, Q. Chen, and Z. Lu, "Genegpt: Augmenting large language models with domain tools for improved access to biomedical information," *Bioinformatics*, vol. 40, no. 2, p. btae075, 2024.

[9] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval," *arXiv preprint arXiv:2401.03985*, 2024.

[10] T. U. Consortium, "Uniprot: the universal protein knowledgebase in 2025," https://www.uniprot.org/, 2025, accessed: 2025-04.

VIII. CONTRIBUTIONS

Dipesh Tripathi: Worked on Approach 4 and documentation.

Ujjwal Bhatta: Worked on Approach 3 and documentation.

Sumaly Bajracharya: Worked on Approach 1 and documentation.

Jeevan Kaphle: Worked on Approach 2 and literature review.