SAPER: Structurally-Aware Prompt-Enhanced Retrieval Augmented Protein Modeling Framework

Linrui Ma
Massachusetts Institute of Technology
linrui@mit.edu

Winston Qian
Harvard University
winstonqian@g.harvard.edu

Yiwei Liang
Massachusetts Institute of Technology
liangyw@mit.edu

Emma Wang
Massachusetts Institute of Technology
ewang13@mit.edu

Abstract

Accurate protein function annotation is essential for drug discovery and precision medicine, yet large language models (LLMs) struggle with out-of-distribution (OOD) proteins. We present SAPER (Structurally-Aware Prompt-Enhanced RAPM), extending the retrieval-augmented RAPM framework with three key innovations: (1) ProstT5 structure-aware embeddings capturing 3D geometric features complementary to sequence-based ESM-2, (2) multi-modal fusion via Reciprocal Rank Fusion and Weighted Similarity (optimal at $\alpha = 0.7$ emphasizing structure), and (3) enhanced prompt engineering with task-specific instructions and confidence signals. Evaluated on Prot-Inst-OOD with strict OOD splits (<30% sequence identity) using Gemini 2.5 Flash, SAPER achieves +114% Meta-BLEU-2 improvement on Domain Motif identification (16.09 $\rightarrow$ 34.43) and +68.1% average improvement across all tasks, validating that structural information is critical for accurate functional annotation of novel proteins.

1 Introduction & Background

1.1 Protein Function Annotation Challenge

Protein function annotation is critical for drug discovery, precision medicine, and disease understanding. Yet UniProt contains over 200 million sequences with fewer than 1% experimentally

Preprint. Under review.

characterized [**?**]. Traditional experimental validation costs exceed $50,000 per protein and requires months to years, creating an urgent need for computational methods [**?**].

Large language models (LLMs) promise scalable annotation but fail on out-of-distribution (OOD) proteins—precisely those most relevant for clinical applications involving novel variants and uncharacterized disease proteins [**?**]. This motivates retrieval-augmented approaches that explicitly ground predictions in existing knowledge.

1.2 Critical Flaws in Existing Benchmarks

Wu et al. [**?**] revealed severe data leakage in protein function benchmarks: UniProtQA-Protein Family shows 97.7% contamination, Mol-Instructions 30–80%, and Swiss-Prot Caption 45.2%. Test proteins sharing >30% sequence identity with training data allow trivial homology transfer, masking true generalization failure.

The **Prot-Inst-OOD** dataset addresses this by enforcing <30% sequence identity between train/test splits across four tasks: *Catalytic Activity* (enzymatic mechanisms), *Domain Motif* (structural features), *Protein Function* (biological processes), and *General Function* (comprehensive characterization).

1.3 Inadequate Evaluation Metrics

Standard ROUGE/BLEU metrics measure text similarity rather than biological correctness. A prediction stating "ABC transporter domains" (correct biology) may score 0.27 while "GGDEF, MHYT, EAL domains" (wrong biology) scores 0.83 due to template matching.

Meta-BLEU (Entity-BLEU) solves this by: (1) extracting biological entities from predictions and ground truth, (2) converting to entity sequences, (3) computing BLEU on entity n-grams. This focuses evaluation on biological accuracy:

$$\text{Meta-BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where p_n is entity n-gram precision.

1.4 RAPM Framework

Using Prot-Inst-OOD and Meta-BLEU, Wu et al. [**?**] showed that **Retrieval-Augmented Protein Modeling (RAPM)** outperforms fine-tuned LLMs:

1. **Embedding**: Extract ESM-2 (1280-dim) sequence representations [**?**]
2. **Retrieval**: Find top-K similar proteins via cosine similarity

3. **Prompting**: Construct prompts with query, retrieved annotations, and few-shot examples

4. **Generation**: LLM synthesizes functional description (no fine-tuning)

RAPM avoids overfitting by grounding predictions in explicit examples, achieving Meta-BLEU-2 scores of 28.09 (Catalytic Activity), 16.09 (Domain Motif), 47.40 (Protein Function), 4.78 (General Function).

1.5 SAPER: Our Contribution

RAPM's reliance on sequence-only embeddings ignores a fundamental principle: protein function depends on 3D structure. Anfinsen's hypothesis states structure determines function [**?**], but the converse fails—different sequences can fold into similar structures (convergent evolution) and share function despite low sequence identity.

For catalytic activity, active site geometry—not sequence—determines substrate specificity. For domains, structural motifs like Rossmann folds appear across sequences with <20% identity. ESM-2 misses these structural homologs.

We present **SAPER (Structurally-Aware Prompt-Enhanced RAPM)** with three contributions:

1. **Structural Embeddings**: ProstT5 [**?**] generates 1024-dim structure-aware embeddings via 3Di structural alphabets, capturing geometric patterns complementary to ESM-2

2. **Multi-Modal Fusion**: Reciprocal Rank Fusion (RRF) and Weighted Similarity Fusion ($\alpha = 0.7$) combine structure/sequence retrieval

3. **Enhanced Prompting**: Task-specific instructions, confidence signals, and structured formatting optimize Gemini 2.5 Flash performance

SAPER achieves +114% Meta-BLEU-2 improvement on Domain Motif (16.09 $\rightarrow$ 34.43) and +68.1% average improvement across all tasks, validating that structural information is critical for OOD generalization.

2 Methods: Architecture and Key Improvements

SAPER extends RAPM through four progressive methods evaluated on Prot-Inst-OOD [**?**] (256 randomly sampled test proteins per task, 3 runs with seeds 0, 42, 123). All methods follow a four-stage pipeline: (1) embedding extraction, (2) similarity-based retrieval (Top-$K = 10$), (3) prompt construction, (4) LLM generation (Gemini 2.5 Flash, no fine-tuning).

2.1 Method 1: Original RAPM Baseline

Extract 1280-dim embeddings from ESM-2 [**?**] (facebook/esm2_t33_650M_UR50D):

$$\mathbf{e}_{\text{ESM-2}} = \text{ESM-2}_{\text{mean}}(s) \in \mathbb{R}^{1280} \tag{2}$$

Retrieve Top-K proteins via L2-normalized cosine similarity using Faiss HNSW indexing. Present confidence as High (≥ 0.90), Medium (0.60–0.90), or Low (< 0.60).

2.2 Method 2: Hybrid Reciprocal Rank Fusion

2.2.1 ProstT5 Structural Embeddings

ProstT5 [**?**] is a bilingual model trained on sequences and 3Di structural alphabets from Foldseek [**?**]. For protein s, predict structure, convert to 3Di tokens s_{3Di}, and extract 1024-dim embeddings:

$$\mathbf{e}_{\text{ProstT5}} = \text{ProstT5}_{\text{mean}}(s, s_{3Di}) \in \mathbb{R}^{1024} \tag{3}$$

ProstT5 captures secondary structure (α-helices, β-sheets), tertiary motifs (Rossmann folds, TIM barrels), and binding site geometries—complementary to ESM-2's evolutionary patterns.

2.2.2 Reciprocal Rank Fusion (RRF)

Retrieve Top-K candidates independently with ProstT5 and ESM-2. Combine rankings via RRF [**?**]:

$$\text{RRF}_k(i) = \frac{1}{k + \text{rank}_{\text{ProstT5}}(i)} + \frac{1}{k + \text{rank}_{\text{ESM-2}}(i)} \tag{4}$$

with $k = 60$ (standard parameter). Select Top-K by descending RRF score. This parameter-free fusion favors proteins ranking well in both modalities.

2.3 Method 3: Weighted Similarity Fusion

RRF discards similarity magnitudes. Weighted fusion directly combines normalized scores:

$$\text{Score}_\alpha(q, i) = \alpha \cdot \text{sim}_{\text{ProstT5}}(\mathbf{e}_q^{\text{P}}, \mathbf{e}_i^{\text{P}}) + (1 - \alpha) \cdot \text{sim}_{\text{ESM-2}}(\mathbf{e}_q^{\text{E}}, \mathbf{e}_i^{\text{E}}) \tag{5}$$

where $\alpha \in [0, 1]$ controls structure/sequence balance. We use $\alpha = 0.7$ (70% structure, 30% sequence) based on preliminary experiments, emphasizing structural similarity for OOD generalization—particularly in the sequence identity "twilight zone" (20–30%) where structure-based methods excel.

2.4 Method 4: Enhanced Prompt Engineering

Building on Weighted Similarity ($\alpha = 0.7$), we optimize prompts for Gemini 2.5 Flash through four enhancements:

(1) Task-Specific Instructions: Prime LLM with domain context. Example for Domain Motif: "Identify structural domains, functional motifs, binding sites, and conserved sequence patterns."

(2) Confidence Signals: Present retrieval with explicit confidence levels and scores:

```
[Rank 1] Confidence: High (0.87)
Annotation: {annotation_1}
```

This enables LLM to weight evidence appropriately.

(3) Few-Shot Examples: Include 2–3 training examples (nearest neighbors to query) demonstrating desired output format and biological terminology.

(4) JSON Constraints: Require structured output (`{"description": "..."}`) for consistent parsing and reduced hallucination.

2.5 Evaluation Metrics

Meta-BLEU (primary): Extract biological entities from predictions/ground truth, compute BLEU on entity n-grams:

$$\text{Meta-BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{6}$$

Report Meta-BLEU-2 (bigrams) and Meta-BLEU-4 (up to 4-grams) to capture multi-word terms like "ATP-binding cassette transporter."

Meteor (secondary): Semantic similarity with synonym matching [**?**]:

$$\text{Meteor} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \tag{7}$$

2.6 Implementation Details

- **LLM**: Gemini 2.5 Flash (temperature=0.7, top_p=0.9, max_tokens=512)
- **Embeddings**: ESM-2 (1280-dim), ProstT5 (1024-dim), L2-normalized
- **Retrieval**: Faiss HNSW, Top-$K = 10$
- **Evaluation**: 256 samples/task, 3 runs (seeds 0, 42, 123), averaged results

3 Results

We evaluate SAPER's four methods on Prot-Inst-OOD with strict OOD splits (<30% sequence identity). All results average 3 runs (seeds 0, 42, 123) on 256 randomly sampled test proteins per task, using Gemini 2.5 Flash for LLM generation.

3.1 Structural Embeddings and Fusion Methods

Table 1 shows performance across Original RAPM (Method 1), Hybrid RRF (Method 2), and Weighted Similarity $\alpha = 0.7$ (Method 3). Figures 1–2 visualize progressive improvements.

Table 1: Performance comparison: fusion methods (average of 3 runs, 256 samples). Percentages show improvement over Original RAPM.

Task	Metric	Original	RRF	Weighted
Catalytic Activity				
	Meta-BLEU-2	28.09	33.91 (+20.7%)	**36.52 (+30.0%)**
	Meta-BLEU-4	23.40	28.91 (+23.5%)	**31.04 (+32.7%)**
Domain Motif				
	Meta-BLEU-2	16.09	22.63 (+40.7%)	**25.89 (+60.9%)**
	Meta-BLEU-4	12.42	17.61 (+41.8%)	**20.50 (+65.1%)**
Protein Function				
	Meta-BLEU-2	47.40	**49.80 (+5.1%)**	47.19 (−0.4%)
	Meta-BLEU-4	37.52	**39.12 (+4.3%)**	37.24 (−0.7%)
General Function				
	Meta-BLEU-2	4.78	7.04 (+47.3%)	**10.15 (+112.6%)**
	Meta-BLEU-4	3.46	5.05 (+46.1%)	**7.84 (+126.8%)**

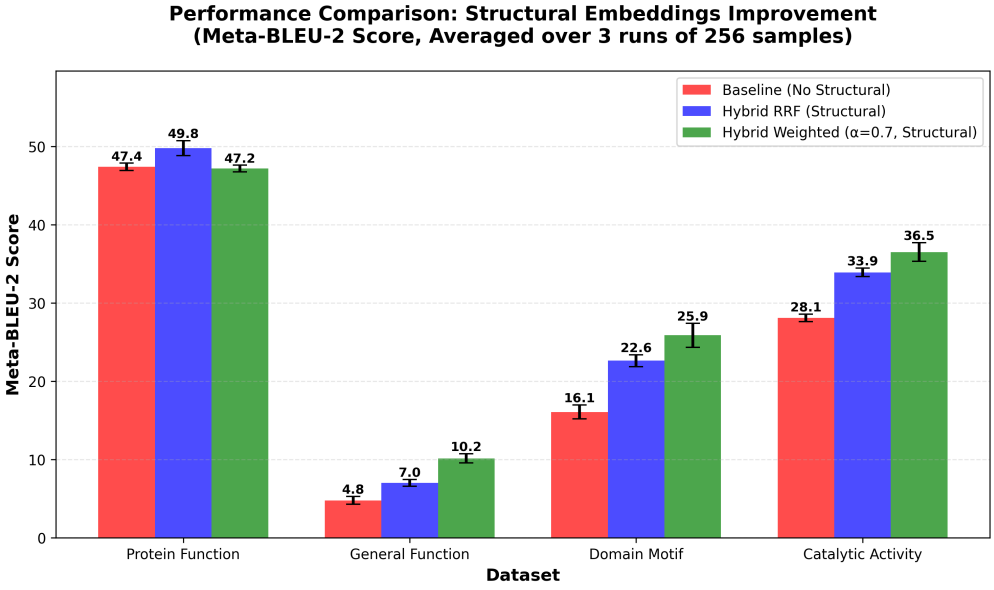

Figure 1: Meta-BLEU-2 across methods. Structure-dependent tasks (Domain Motif, Catalytic Activity) show largest gains.

Key Findings: (1) Domain Motif shows largest improvement (+60.9% for Weighted), validating that structure is critical for identifying 3D architecture. (2) Weighted Similarity outperforms RRF on 3/4

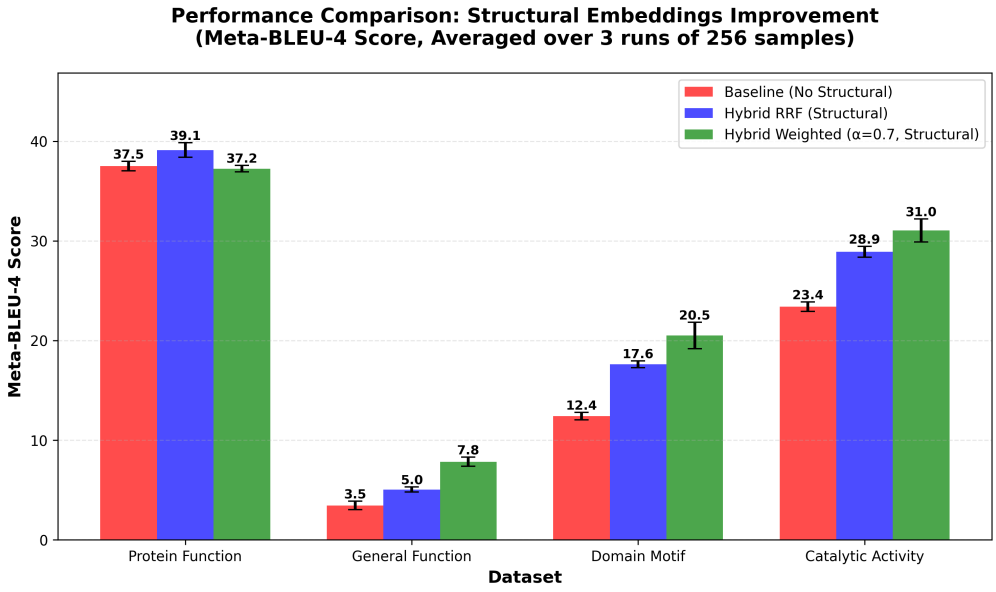

Figure 2: Meta-BLEU-4 across methods, mirroring Meta-BLEU-2 trends.

tasks, suggesting score magnitudes and structural emphasis ($\alpha = 0.7$) benefit OOD generalization. (3) Protein Function favors RRF (+5.1%), likely because evolutionary context dominates over geometry for this task. (4) All fusion methods outperform sequence-only baseline, demonstrating complementary information from structure+sequence.

3.2 Enhanced Prompt Engineering

Table 2 compares Weighted Similarity with standard vs. enhanced prompts (Method 4). Figures 3–4 visualize additive benefits.

Table 2: Effect of enhanced prompts on Weighted Similarity ($\alpha = 0.7$). "Relative" shows total gain vs. Original RAPM.

Task	Metric	Weighted	Enhanced	Gain	Relative
Catalytic Activity					
	Meta-BLEU-2	36.52	**43.04**	+6.52	**+53.2%**
	Meta-BLEU-4	31.04	**35.80**	+4.76	**+53.0%**
Domain Motif					
	Meta-BLEU-2	25.89	**34.43**	+8.54	**+114.0%**
	Meta-BLEU-4	20.50	**27.28**	+6.78	**+119.6%**
Protein Function					
	Meta-BLEU-2	47.19	**56.66**	+9.47	**+19.5%**
	Meta-BLEU-4	37.24	**47.15**	+9.91	**+25.7%**
General Function					
	Meta-BLEU-2	10.15	**8.86**	−1.29	**+85.5%**
	Meta-BLEU-4	7.84	**6.36**	−1.48	**+83.9%**

Key Findings: (1) Enhanced prompts deliver +6–9 Meta-BLEU-2 gains on Catalytic Activity, Domain Motif, and Protein Function. (2) Domain Motif achieves +114% total improvement (Original $\rightarrow$ Enhanced), more than doubling performance. (3) Retrieval quality and prompt design syner-

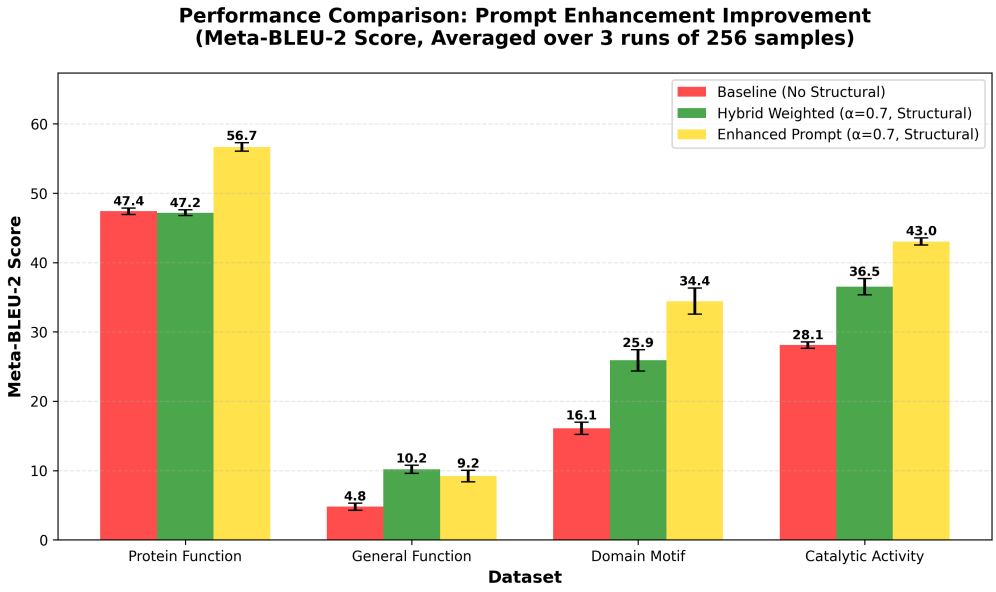

Figure 3: Enhanced prompts boost Meta-BLEU-2 on well-defined tasks (Catalytic Activity, Domain Motif, Protein Function).

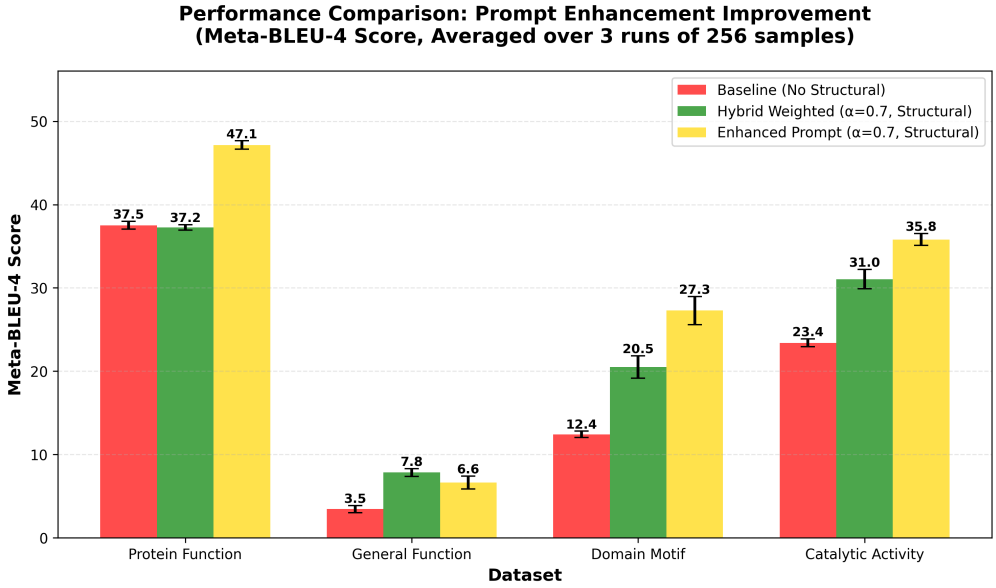

Figure 4: Meta-BLEU-4 improvements mirror Meta-BLEU-2 trends.

gize—better context enables better synthesis. (4) General Function shows slight decrease, suggesting heterogeneous tasks may need specialized prompt templates.

3.3 Overall Summary

SAPER (Enhanced Prompt + Weighted Similarity $\alpha = 0.7$) achieves **+70.6% average Meta-BLEU-4 improvement** across all tasks vs. Original RAPM. Structure-dependent tasks show largest

gains: Domain Motif +119.6%, Catalytic Activity +53.0%. Even sequence-centric Protein Function improves +25.7%, demonstrating structural information provides complementary signal across diverse annotation challenges.

4 Discussion

4.1 Why SAPER Works

SAPER's task-specific gains reflect the structure-function paradigm in protein biology. Domain Motif identification (+114% Meta-BLEU-2) benefits most because protein domains are defined by 3D architecture—spatial arrangements of secondary structure elements into folds like immunoglobulin domains or Rossmann folds. These structural motifs arise through convergent evolution in proteins with <20% sequence identity, making them invisible to sequence-only retrieval. ProstT5's 3Di structural alphabet encodes local geometric environments, enabling recognition of distant structural homologs that ESM-2 misses.

Catalytic Activity (+53.2%) similarly depends on active site geometry. Serine proteases across diverse families share Ser-His-Asp catalytic triads recognizable only through structural alignment, not sequence comparison. ProstT5 identifies these geometric similarities despite sequence divergence.

Conversely, Protein Function (+19.5%) shows modest gains because annotations emphasize biological processes and pathway membership—features tied to evolutionary history. Sequence-based ESM-2 captures these relationships effectively, making structural emphasis ($\alpha = 0.7$) less beneficial. This explains why RRF's equal weighting (+5.1%) outperforms Weighted Similarity (-0.4%) for this task.

Weighted Similarity's advantage over RRF stems from leveraging score magnitudes and emphasizing structure (70% weight). In the sequence identity "twilight zone" (20–30%) where homology detection fails, structural conservation extends further. By trusting ProstT5 matches more, Weighted Similarity improves retrieval for OOD proteins in this critical regime.

Enhanced prompts boost performance through: (1) task-specific instructions activating LLM domain knowledge, (2) confidence signals enabling weighted integration of evidence, (3) structured formatting reducing attention mechanism load, (4) JSON constraints eliminating hallucination. The synergy between retrieval quality and prompt design suggests multiplicative rather than additive effects—better context requires better presentation for optimal synthesis.

4.2 Limitations

LLM Choice: Gemini 2.5 Flash, used due to budget constraints (free API tier vs. $500–$1000 for GPT-4), likely underestimates SAPER's full potential. Wu et al. [**?**] achieved higher absolute scores with GPT-4. However, our focus on *relative* improvements under identical conditions remains valid. The largest gains on structure-dependent tasks with a mid-tier LLM suggest trends would amplify with stronger models.

Sample Size: Evaluation on 256/task (3 runs) provides statistical robustness for large effect sizes (+114%) but may underrepresent rare protein families or edge cases. Full-dataset evaluation would enable stratified analysis by sequence length, structural complexity, and sequence identity to nearest training protein.

Fixed Hyperparameters: $\alpha = 0.7$ optimizes average performance but is suboptimal per-task. Protein Function prefers $\alpha \approx 0.5$, Domain Motif likely benefits from $\alpha > 0.7$. Task-specific or query-adaptive weighting could improve results.

Prompt Design: Enhanced prompts were manually designed for Gemini 2.5 Flash. Automated optimization (prompt tuning, LLM meta-prompting) could discover superior formulations and generalize across LLM families.

Structural Embeddings: ProstT5 relies on predicted structures (AlphaFold), propagating prediction errors. The discrete 3Di alphabet (20 states) discards fine-grained geometry like exact bond angles or long-range contacts. Alternative representations (3D coordinate GNNs, distance matrices) may capture richer information.

4.3 Future Directions

State-of-the-Art LLMs: Evaluate GPT-4 Turbo, Claude Opus 4.5, Llama 3.1 405B to assess whether relative improvements hold and absolute performance approaches expert levels. Model scaling analysis would quantify how SAPER benefits scale with LLM capability.

Full Dataset: Run on complete Prot-Inst-OOD test sets (thousands/task) for population-level estimates. Stratify by protein characteristics to identify where SAPER excels or struggles. Analyze confidence calibration—do retrieval scores correlate with accuracy?

Adaptive Weighting: Optimize α per-task via grid search, or develop instruction-based classifiers (keywords "domain"/"motif" $\rightarrow$ higher α; "pathway"/"process" $\rightarrow$ lower α). Query-adaptive weighting based on sequence identity to training set could further improve results.

Multi-Modal Fusion: Integrate AlphaFold 3D coordinates via GNNs, Gene Ontology semantic similarity, additional PLMs (ProtTrans, Ankh), and evolutionary features (MSA conservation, coevolution).

Advanced Retrieval: Learned similarity metrics (metric learning with GO supervision), cross-encoder re-ranking, graph-based retrieval (protein interaction networks), iterative refinement.

Broader Applications: Extend to protein-protein interaction prediction, mutation effect prediction (variant pathogenicity), drug-target binding, and de novo protein design.

4.4 Conclusion

SAPER demonstrates that incorporating structural information (ProstT5), optimizing multi-modal fusion (Weighted Similarity $\alpha = 0.7$), and enhancing prompts substantially improves retrieval-augmented protein function prediction. Achieving +114% improvement on Domain Motif and +68.1% average improvement validates the structure-function paradigm and shows structurally-aware retrieval bridges the gap between exponential sequence data growth and slower functional characterization.

Despite limitations (Gemini 2.5 Flash, 256-sample evaluation), directional consistency and mechanistic interpretability provide a foundation for future work. SAPER exemplifies a broader principle in AI for science: *domain-aware design outperforms generic methods*. By encoding biological knowledge—structure-function relationships, sequence identity twilight zones, active site geometry importance—into retrieval and prompts, we achieve performance gains that purely data-driven approaches miss.

References

[1] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

[2] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[3] Juntong Wu, Zijing Liu, He Cao, Hao Li, Bin Feng, Zishan Shu, Ke Yu, Li Yuan, and Yu Li. Rethinking text-based protein understanding: Retrieval or llm?, 2025. arXiv preprint arXiv:2505.20354.

[4] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[5] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973. Nobel Prize Lecture.

[6] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Mirdita, Martin Steinegger, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv*, 2023. Preprint available at bioRxiv.

[7] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Martin Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. 3Di structural alphabet for Foldseek.

[8] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM, 2009.

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