LLM Agent Based Protein Function Prediction

Fernando Zhapa-Camacho, Olga Mashkova, Robert Hoehndorf[†], and Maxat Kulmanov
Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology,
Thuwal, Makkah Region, Saudi Arabia
[†]*E-mail: robert.hoehndorf@kaust.edu.sa*
www.kaust.edu.sa

Protein function prediction remains a fundamental challenge in computational biology. Here, we present a Large Language Model (LLM) agent-based system that improves protein function prediction performance using knowledge-augmented reasoning and multi-source evidence synthesis.

Our approach integrates computational predictions with structured protein metadata, scientific literature, and ontological knowledge through a multi-stage reasoning process. An LLM agent equipped with specialized tools progressively refines functional predictions by querying constraints, cross-referencing evidence, and ensuring biological plausibility. Furthermore, the system provides detailed explanations for each prediction update, documenting the reasoning process and evidence sources.

We evaluate our approach against established baseline methods across three Gene Ontology sub-ontologies using four complementary metrics, achieving superior performance in threshold-dependent measures, attaining the lowest $S_{\min}$ scores across all ontologies and the best $F_{\max}$ for Molecular Function and Cellular Component ontologies. We make our code publicly available at `https://github.com/bio-ontology-research-group/go-agent`.

Keywords: Gene Ontology; Protein Function Prediction; Large Language Models; Agents.

1. Introduction

Protein function prediction is one of the key challenges in modern biology and bioinformatics as it enables better understanding of the roles and interactions of proteins within living systems. Accurate functional descriptions of proteins are necessary for tasks such as identification of drug targets, understanding disease mechanisms, and improving biotechnological applications in industry. While predicting protein structures has become increasingly accurate in recent years,[1] predicting protein function remains challenging due to the small number of known functions combined with their complexity and interactions.

Functions of proteins are described using the Gene Ontology (GO).[2] GO includes three sub-ontologies for describing molecular functions (MFO) of a single protein, biological processes (BPO) to which proteins can contribute, and cellular components (CCO) where proteins are active. Researchers identify protein functions based on both targeted experiments and high-throughput experimental approaches, generating scientific reports which are then taken by database curators and added to knowledge bases. Currently, the UniProtKB/Swiss-Prot database[3] contains reviewed GO annotations for thousands of organisms and more than 570,000 proteins.

Protein function prediction methods can rely on different sources of information such as sequence, interactions, protein tertiary structure, literature, or the information provided in GO.[4–16] The methods may use sequence domain annotations,[5,6,8,11,17] directly apply deep convolutional neural networks (CNN)[13] or language models such as LSTMs[9] and transformers,[14] or use pretrained protein language models[10,15] to represent amino acid sequences. Models may also incorporate protein–protein interactions through knowledge graph embeddings,[12,16] approaches using k-nearest neighbors,[17] and graph convolutional neural networks.[6] Also, natural language models applied to scientific literature have been successful in automated function prediction.[8]

Despite these advances, current protein function prediction methods face several limitations. Most approaches treat prediction as a static classification problem without incorporating the rapidly growing nature of biological literature or providing multi-step reasoning explanations for their predictions. While some methods offer interpretability through attention mechanisms[14] or graph-based explanations in GNNs[18] or LLMs[19] they typically highlight important features rather than providing structured reasoning that traces through multiple evidence sources and biological constraints. Traditional methods, although capable of integrating multiple data types as demonstrated by approaches like GOLabeler,[5] often rely on fixed integration schemes that cannot re-evaluate evidence based on context-specific biological knowledge. These methods cannot easily incorporate new biological insights without retraining or adapt their integration strategy based on the reliability of different evidence sources for specific protein families or functional categories. Furthermore, while these methods provide confidence scores, they lack the ability to generate natural language explanations that trace the logical steps from evidence to conclusion, making it difficult for biologists to assess the reliability of predictions or understand the biological reasoning supporting specific functional assignments.

The emergence of large language models (LLMs) has introduced new capabilities for integrating heterogeneous information sources and performing evidence-based reasoning in computational biology. Traditional machine learning approaches for protein function prediction rely on fixed feature representations and statistical patterns learned from training data, limiting their ability to incorporate new biological knowledge or provide explanations for their predictions. In contrast, LLM agent-based systems, exemplified by frameworks such as LangChain[a] and multi- agent systems like CAMEL-AI,[20] can access external knowledge bases on-demand, synthesize information from scientific literature, and apply domain-specific constraints through specialized tools. The key advantage of LLM agents over standalone LLMs lies in their ability to perform multi-step reasoning processes, where intermediate results can be validated against external knowledge sources and refined through iterative consultation of different information repositories. This agent-based approach enables the combination of computational predictions

[a] https://github.com/langchain-ai/langchain

with structured biological knowledge and experimental evidence from literature, potentially improving both the accuracy and interpretability of protein function predictions.

In this work, we propose an LLM agent-based system that addresses these limitations by combining computational predictions with LLM-generated reasoning and evidence synthesis. Our approach integrates multiple information sources including supervised predictions from a multi-layer perception (MLP) network trained on SwissProt annotations, homology-based predictions from DiamondScore methods, structured protein metadata from UniProtKB, relevant scientific literature from PubMed[b], and ontological knowledge from the Gene Ontology database. The system employs a multi-stage reasoning process where an LLM agent equipped with specialized tools progressively refines functional predictions through evaluation of available evidence and ontological consistency checking. Unlike traditional black-box approaches, our method provides detailed explanations for each prediction update, tracing the reasoning process and evidence sources that influenced the final annotations. We evaluate our approach against established baseline methods and demonstrate improved performance in semantic accuracy and optimal threshold-dependent metrics, while providing interpretable insights that support biological hypothesis generation and validation.

2. Materials and Methods

2.1. *Gene Ontology and Dataset*

We obtained all proteins that were manually curated and reviewed from the UniProtKB/Swiss-Prot Knowledgebase. We generated a time-based test set by following the CAFA[21] challenge time-based approach. For training, we took all proteins from release v2023_05 (08-Nov-2023).[3] For validation, we took the proteins that collected annotations in the release v2024_06 (27-Nov-2024). For testing, we took the proteins that appeared in the release v2025_03 (18-Jun-2025). We selected time based evaluation with these dates to make sure that the LLMs we use were not trained on our testing set. We filtered proteins with experimental functional annotations using evidence codes EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, HTP, HDA, HMP, HGI, HEP. The dataset contains $84,748$ reviewed and manually annotated proteins and a subset of filtered Gene Ontology annotations. The dataset includes comprehensive annotations from multiple sources: UniProtKB entries containing protein metadata and functional descriptions, PubMed abstracts associated with each protein through database cross-references, InterPro domain annotations obtained via InterProScan, and curated Gene Ontology annotations across all three sub-ontologies (Molecular Function, Biological Process, and Cellular Component). Table 1 summarizes the datasets for each sub-ontology.

We use Gene Ontology (GO)[2] version released on 09-Oct-2023. In addition, we use `go-computed-taxon-constraints` file to query taxon constraints axioms.

2.2. *Baseline comparison methods*

To evaluate our LLM agent-based approach for protein function prediction, we compare against several established methods that provide diverse perspectives on computational function pre-

[b] https://pubmed.ncbi.nlm.nih.gov/

Table 1. Summary of the UniProtKB-SwissProt dataset. The table shows the number of GO terms, total number of proteins, number of groups of similar proteins, number of proteins in training, validation and testing sets for the UniProtKB-SwissProt dataset.

Ontology	Terms	Proteins	Training	Validation	Testing
MFO	7,168	46,573	45,365	801	407
BPO	20,848	62,613	60,615	1,225	773
CCO	2,905	61,591	59,926	1,126	539

diction.

Given the inherent class imbalance in Gene Ontology (GO) annotations and the hierarchical structure imposed by the true-path rule, certain GO terms appear more frequently than others in training data. The naive classifier, introduced by the Critical Assessment of Functional Annotation (CAFA) challenge,[22] exploits this frequency distribution by assigning GO terms to all proteins based solely on their prevalence in the training set. For each query protein p and GO term f, the prediction score is computed as $S(p, f) = \frac{N_f}{N_{total}}$, where N_f represents the number of training proteins annotated with GO term f, and N_{total} denotes the total number of training proteins. This baseline provides a lower bound for performance evaluation and helps assess whether more sophisticated methods provide meaningful improvements over frequency-based predictions.

The DiamondScore method is based on the sequence similarity score obtained by Diamond.[23] The method aims to find similar sequences from the training set and transfer their annotations using the normalized bitscore to compute the prediction score for a query sequence q: $S(q, f) = \frac{\sum_{s \in E} bitscore(q,s) * I(f \in T_s)}{\sum_{s \in E} bitscore(q,s)}$, where E is a set of similar sequences filtered by e-value of 0.001, T_s is a set of true annotations of a protein with sequence s, and I is an indicator function which returns 1 if the condition is true and 0 otherwise.

MLP baseline uses ESM2[24] embeddings as input features. The network architecture consists of two MLP blocks with residual connections, followed by a sigmoid classification layer. Each MLP block applies the transformation $\text{MLPBlock}(\mathbf{x}) = \text{Dropout}(\text{BatchNorm}(\text{ReLU}(W\mathbf{x} + b)))$. The first MLP block reduces the input dimensionality from 2,560 to 1,024, the second block maintains the 1,024-dimensional representation while incorporating a residual connection, and finally, the classification layer produces predictions using sigmoid activation where the output dimensionality matches the number of classes in each GO sub-ontology. We train separate models for each sub-ontology (Biological Process, Molecular Function, and Cellular Component).

DeepGO-PLUS[13] integrates sequence-based predictions from a Convolutional Neural Network (CNN) model with similarity-based DiamondScore predictions. The CNN model employs a one-dimensional convolutional neural network to identify sequence motifs associated with specific GO functions directly from amino acid sequences.

DeepGraphGO[6] integrates sequence-derived features with protein-protein interaction (PPI) network information using graph convolutional neural networks. The method combines

InterPro domain annotations with topological information from PPI networks, allowing the model to leverage both intrinsic protein properties and interaction context for function prediction. The graph convolutional architecture enables the propagation of functional information through the protein interaction network, potentially capturing functional relationships that emerge from protein complex formation and pathway participation. We implemented DeepGraphGO based on the original manuscript specifications and trained the model using our standardized dataset for fair comparison.

SPROF-GO[25] uses the ProtT5-XL-U50 protein language model to extract protein sequence embeddings and learns an attention-based neural network model. The model incorporates the hierarchical structure of GO into the neural network and predicts functions that are consistent with hierarchical relations of GO classes. Furthermore, SPROF-GO combines sequence similarity-based predictions using a homology-based label diffusion algorithm. We used the trained models (v1) provided by the authors to evaluate them on the time-based dataset.

2.3. *LLM-based agent system*

The LLM agent system is implemented using the CAMEL-AI[20] framework, with the Gemini-Flash-2.0 and GPT-4.1 nano language models as the underlying LLM engine, which were selected for their strong performance on scientific reasoning tasks and multi-step problem solving. The agent operates with a context window that accommodates protein-specific information including UniProtKB metadata, relevant literature abstracts, and specialized tool outputs. Both Gemini-Flash-2.0 and GPT-4.1 nano have knowledge-cutoff date of June 2024 which is before the release date of our test protein annotations.

The implemented LLM agent-based system integrates multiple information sources and computational predictions to refine protein function annotations through structured reasoning and evidence synthesis. The system operates as a specialized GO annotation curator that processes individual proteins through a multi-stage refinement pipeline following a systematic workflow. The system aim is to refine scores for already known functions rather than discovering new ones.

For each target protein, the system begins by assembling a comprehensive information profile that includes initial computational predictions from MLP models trained on SwissProt annotations, homology-based predictions from DiamondScore sequence similarity methods, structured protein metadata and functional descriptions from UniProtKB entries, relevant scientific literature abstracts from PubMed associated with the protein, InterPro domain annotations obtained through InterProScan, and ontological knowledge with taxonomic constraints from the Gene Ontology database (Figure 1 section (a)). The agent operates with a contextual system message that establishes its role as a GO annotation curator for the specific protein, incorporating UniProtKB functional information and relevant literature abstracts into its reasoning context to ensure responses are grounded in protein-specific biological knowledge.

The LLM agent is equipped with four specialized tools that enable adaptive information retrieval and prediction updates during the reasoning process. The InterPro tool retrieves and processes InterPro domain annotations, mapping them to associated GO terms through the InterPro2GO mappings. The GO tool provides detailed information about specific GO terms, including their definitions, hierarchical relationships, and associated prediction scores. The TaxonConstraints tool queries organism-specific constraints from the Gene Ontology, identifying GO terms that are taxonomically restricted or prohibited for the target organism. The Update tool allows the agent to modify prediction scores based on its analysis.

The prediction refinement operates through a two-stage prompting strategy that follows the agent's analytical workflow. In the first stage, the agent receives all GO terms with initial prediction scores ≥ 0.1 along with their associated information including prediction scores, DiamondScore homology evidence, and annotation frequencies in the training data. The agent analyzes each term considering multiple evidence sources: annotation frequency patterns to identify potentially underrepresented terms with frequencies below 200, supporting evidence from InterPro domains and sequence similarity, and resolution of conflicts between different evidence sources. The system integrates multiple types of biological constraints during this analysis, using taxonomic constraints to ensure predicted functions are biologically plausible for the target organism, while annotation frequency analysis helps identify potentially novel or underrepresented functions that may be missed by frequency-biased training data (Figure 1 section (b)).

In the second stage, the agent applies its analysis to update GO term scores with specific constraints: changes are limited to increments or decrements of maximum x to minimize excessive modifications, and updates must be accompanied by detailed rationales explaining the evidence and reasoning behind each change. We empirically determined the value of x to be 0.2 for the current experiment using a limited random search. The agent provides structured output including current versus recommended scores, confidence levels, and resolution of conflicting evidence. Throughout this process, the agent synthesizes evidence from sequence homology (DiamondScore), structural domains (InterPro), experimental literature (PubMed abstracts), and ontological relationships (GO hierarchy) to make informed decisions about functional assignments (Figure 1 section (c)).

The iterative nature of this workflow allows the agent to reconsider initial predictions in light of multiple evidence sources, potentially identifying functions that were initially scored low due to training data biases but are well-supported by domain annotations or literature evidence. This approach enables the system to provide not only refined predictions but also detailed explanations that trace the reasoning process and evidence sources that influenced each functional assignment, supporting biological hypothesis generation and enabling researchers to understand the rationale behind functional annotations. We show an example of the agent creation and reasoning process in our Supplementary Material[c].

3. Results

3.1. *Evaluation Metrics*

We evaluate the performance of our protein function prediction models using four complementary metrics established by the CAFA challenge:[22] three protein-centric measures ($F_{\max}$, $S_{\min}$,

[c] `https://github.com/bio-ontology-research-group/go-agent/blob/main/supplementary.pdf`

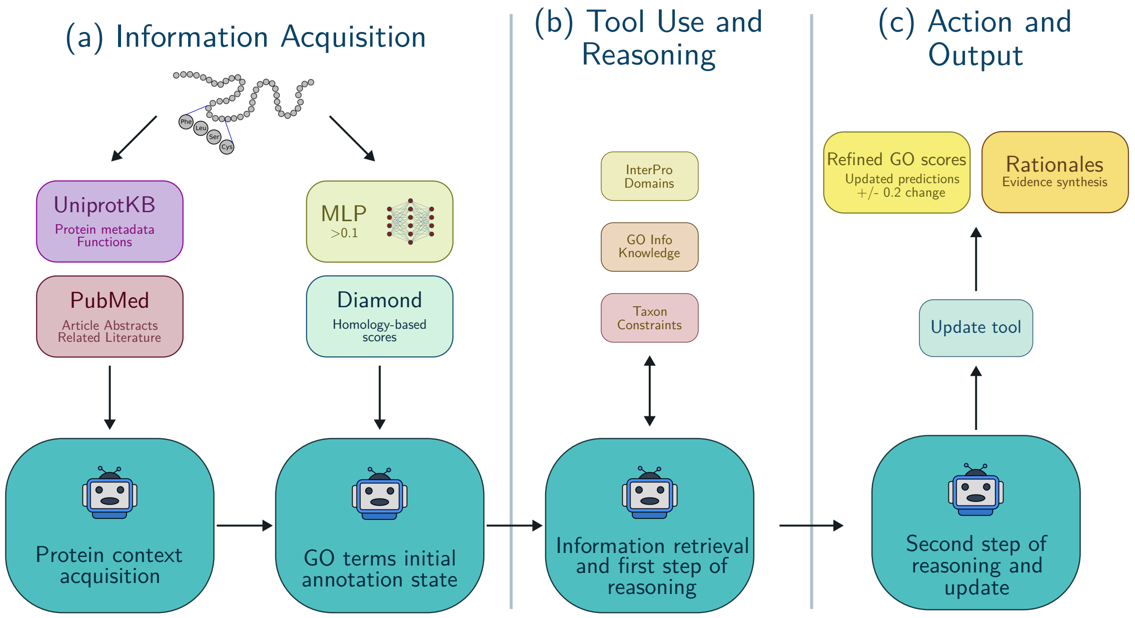

Fig. 1. LLM Agent-Based Protein Function Prediction Workflow. (a) Agent acquires protein metadata information from UniprotKB and literature information from PubMed. Furthermore, the Agent also acquires GO prediction scores from a MLP and Diamond. (b) Agent performs tool calls, integrates information and performs a reasoning step. (c) Agent produces refined GO scores with explanations.

and AUPR) and one class-centric measure (AUC). F_{max} represents the maximum protein-centric F-measure achieved across all possible prediction thresholds, balancing precision and recall by identifying the optimal threshold that maximizes their harmonic mean. S_{min} quantifies the semantic distance between predicted and true annotations by incorporating the information content of GO terms, accounting for the hierarchical structure of the Gene Ontology and penalizing errors based on term specificity. AUPR measures the area under the precision-recall curve, providing a threshold-independent comprehensive view of model performance across all operating points. AUC represents a class-centric evaluation where we compute the area under the ROC curve for each individual GO class and calculate the average across all classes, assessing the model's ability to distinguish between positive and negative instances for each functional category independently. These metrics provide comprehensive assessment across different aspects of prediction quality, capturing both threshold-dependent and threshold-independent performance characteristics.

3.2. *Evaluation*

The results presented in Table 2 demonstrate the comparative performance of our LLM agent-based approach (Gemini-Flash-2.0 and GPT-4.1 nano) against established baseline methods across three Gene Ontology sub-ontologies: Molecular Function Ontology (MFO), Biological Process Ontology (BPO), and Cellular Component Ontology (CCO).

Both LLM-based approaches achieve superior performance across most evaluation metrics, with complementary strengths. Gemini-Flash-2.0 demonstrates the strongest performance in threshold-dependent measures, achieving the highest $F_{\max}$ scores for MFO (0.718) and CCO (0.737), and the best semantic accuracy with $S_{\min}$ values of 5.935 for MFO, 25.381 for BPO and 6.190 for CCO. GPT-4.1 nano achieves the second best performance for $F_{\max}$ and $S_{\min}$ for MFO and CCO, without sacrificing threshold-independent metrics such as AUC. Table 2 contains results for GPT-2.1 nano and Gemini-Flash 2.0 averaged across 5 versions of each experiment and Table 3 integrates both mean and standard deviation results. Among the computational baselines, the MLP+DiamondScore combination demonstrates the strongest overall performance, serving as the most competitive non-LLM approach. This hybrid method achieves best results AUPR and AUC scores across all subontologies. DeepGOPlus shows strong performance particularly in threshold- independent metrics, while methods like DeepGraphGO and SPROF-GO exhibit lower performance across most measures. The naive baseline, as expected, provides the lowest performance bounds across all metrics.

The results indicate that incorporating LLM-based reasoning with domain knowledge and literature evidence can improve protein function prediction, particularly when optimizing for balanced precision-recall performance and semantic accuracy.

Table 2. The comparison of performance on the time-based dataset. For our method, we report the mean metrics accross 5 runs of experiments. Best performing results are bold.

Method	$F_{\max}$			$S_{\min}$			AUPR			AUC		
	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
Naive	0.266	0.262	0.568	12.452	30.865	9.822	0.126	0.146	0.410	0.500	0.500	0.500
DiamondScore (DS)	0.699	0.452	0.689	6.262	25.907	6.488	0.568	0.320	0.492	0.877	0.727	0.826
DeepGOPlus	0.695	0.454	0.719	6.313	25.873	6.456	0.688	0.382	0.746	0.931	0.833	0.906
SPROF-GO	0.649	0.398	0.681	7.988	27.278	7.948	0.628	0.333	0.695	0.746	0.692	0.749
DeepGraphGO	0.611	0.378	0.651	8.390	28.118	7.997	0.546	0.295	0.620	0.781	0.696	0.791
MLP	0.642	0.414	0.693	7.364	27.440	7.530	0.642	0.354	0.723	0.957	0.868	0.936
MLP+DS	0.706	**0.463**	0.718	6.177	25.495	6.365	**0.712**	**0.402**	**0.755**	**0.964**	**0.874**	**0.949**
GPT-4.1 nano	0.709	0.462	0.722	6.132	25.515	6.332	0.704	0.400	0.748	**0.964**	**0.874**	**0.949**
Gemini-Flash 2.0	**0.718**	0.460	**0.737**	**5.935**	**25.381**	**6.190**	0.695	0.386	0.728	0.945	0.867	0.946

Table 3. Performance variance of LLM-based systems. We setup LLM temperature at 0.3. We report the mean and standard deviation across 5 runs of experiments.

Model	Ontology	$F_{\max}$	$S_{\min}$	AUPR	AUC
GPT	MF	0.709 ± 0.003	6.132 ± 0.046	0.704 ± 0.005	0.964 ± 0.000
	BP	0.462 ± 0.001	25.515 ± 0.018	0.400 ± 0.001	0.874 ± 0.000
	CC	0.722 ± 0.002	6.332 ± 0.053	0.748 ± 0.002	0.949 ± 0.000
Gemini	MF	0.718 ± 0.002	5.935 ± 0.099	0.695 ± 0.003	0.945 ± 0.003
	BP	0.460 ± 0.003	25.381 ± 0.111	0.386 ± 0.001	0.867 ± 0.003
	CC	0.737 ± 0.005	6.190 ± 0.052	0.728 ± 0.007	0.946 ± 0.002

We additionally evaluate the impact of the external information provided to the agent:

UniProtKB metadata and PubMed literature information. The results in Table 4 demonstrate that the integration of external information sources has a nuanced impact on performance across different GO ontologies and evaluation metrics. For both GPT and Gemini models, the inclusion of metadata and abstracts shows improvements in $F_{\max}$ and $S_{\min}$ scores across most ontologies, with the most gains observed in CCO predictions. Interestingly, the AUPR and AUC metrics show mixed results, with some configurations performing better without external information, suggesting that the benefits of incorporating metadata and literature abstracts are context-dependent and metric-specific. This indicates that while external information sources can enhance certain aspects of GO term prediction, their integration requires careful consideration of the specific evaluation criteria and ontology domains being targeted.

Table 4. Results for the ablation study of the initial context for GPT-4.1 nano and Gemini-Flash-2.0. Best performing results are bold.

Model	**Method**	$F_{\max}$			$S_{\min}$			AUPR			AUC		
		MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
GPT	Without Metadata	0.705	0.461	0.720	6.223	25.537	6.414	0.704	0.399	0.738	0.957	**0.874**	**0.949**
	Without Abstracts	0.708	0.461	**0.722**	**6.062**	25.530	6.351	**0.707**	**0.400**	0.746	0.961	**0.874**	**0.949**
	All info	**0.709**	**0.462**	**0.722**	6.132	**25.515**	**6.332**	0.704	**0.400**	**0.748**	**0.964**	**0.874**	**0.949**
Gemini	Without Metadata	0.710	0.458	0.724	6.129	25.760	6.216	0.700	**0.387**	0.734	**0.952**	0.871	**0.946**
	Without Abstracts	0.717	0.457	0.731	6.118	25.508	6.223	**0.703**	0.385	**0.735**	0.951	0.870	0.944
	All info	**0.718**	**0.460**	**0.737**	**5.935**	**25.381**	**6.190**	0.695	0.386	0.728	0.945	0.867	**0.946**

4. Conclusions and Future work

In this work, we introduced an LLM agent-based system for protein function prediction that effectively integrates computational predictions with domain knowledge and scientific literature through multi-stage reasoning. Our approach addresses key limitations of traditional methods by providing both improved predictive performance and interpretable explanations for functional assignments. The system demonstrates superior performance in threshold-dependent metrics, achieving the highest $F_{\max}$ for Molecular Function and Cellular Component subontologies and and optimal semantic accuracy across all Gene Ontology sub-ontologies. Beyond quantitative improvements, our method provides detailed reasoning traces that document the evidence and constraints considered during prediction refinement, enabling researchers to understand and validate functional assignments. The results establish that knowledge-augmented LLM agents can effectively combine heterogeneous biological information sources, leveraging both computational predictions and domain expertise to advance protein function annotation. This paradigm shift from purely data-driven approaches to reasoning-based systems opens new possibilities for interpretable and evidence-grounded computational biology applications.

While our current approach demonstrates significant improvements in protein function prediction, several limitations present opportunities for future enhancements. Our single-agent architecture, though effective, may benefit from a multi-agent collaborative framework where specialized agents with distinct expertise domains work together to provide more comprehensive functional annotations through agent negotiation and consensus building. A significant limitation is the system's reliance on pre-curated literature abstracts, which could be addressed by incorporating real-time literature search capabilities that allow agents to actively query and retrieve updated relevant publications, patents, and preprints, accessing the most recent research findings beyond manually curated associations. Additionally, our method currently lacks integration of three-dimensional protein structure information from sources like the Protein Data Bank[26] and AlphaFold[1] predictions, which represents a critical knowledge gap given the fundamental relationship between protein structure and function. The system's access to Gene Ontology background knowledge is limited to basic term definitions and taxonomic constraints, and could be expanded to include comprehensive ontological relationships, semantic similarities between GO terms and annotation provenance information. Furthermore, our current approach processes each protein independently, missing potential functional relationships and dependencies that emerge from protein-protein interactions, co-expression patterns, and shared evolutionary history. Future multi-agent systems could incorporate network-based reasoning to consider functional coherence within protein complexes, metabolic pathways, and regulatory modules, leading to more biologically consistent and systems-level functional annotations while maintaining the interpretability and evidence-based reasoning demonstrated in our current approach.

References

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature* **596**, 583 (July 2021).
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, M. J. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. I. Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology, *Nature Genetics* **25**, 25 (May 2000).
3. T. U. Consortium, UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research* **51**, D523 (11 2022).
4. N. Zhou, Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, L. Davis, T. Dogan, V. Atalay, A. S. Rifaioglu, A. Dalkiran, R. Cetin-Atalay, C. Zhang, R. L. Hurto, P. L. Freddolino, Y. Zhang, P. Bhat, F. Supek, J. M. Fernández, B. Gemovic, V. R. Perovic, R. S. Davidović, N. Sumonja, N. Veljkovic, E. Asgari, M. R. Mofrad, G. Profiti, C. Savojardo, P. L. Martelli, R. Casadio, F. Boecker, I. Kahanda, N. Thurlby, A. C. McHardy, A. Renaux, R. Saidi, J. Gough, A. A. Freitas, M. Antczak, F. Fabris, M. N. Wass, J. Hou, J. Cheng, J. Hou, Z. Wang, A. E. Romero, A. Paccanaro, H. Yang, T. Goldberg, C. Zhao, L. Holm, P. Törönen, A. J. Medlar, E. Zosa, I. Borukhov, I. Novikov, A. Wilkins, O. Lichtarge, P.-H. Chi, W.-C. Tseng, M. Linial, P. W. Rose, C. Dessimoz, V. Vidulin, S. Dzeroski, I. Sillitoe, S. Das, J. G. Lees, D. T. Jones, C. Wan, D. Cozzetto, R. Fa, M. Torres, A. W. Vesztrocy, J. M. Rodriguez, M. L. Tress, M. Frasca, M. Notaro, G. Grossi, A. Petrini, M. Re, G. Valentini, M. Mesiti, D. B. Roche, J. Reeb, D. W. Ritchie, S. Aridhi, S. Z. Alborzi, M.-D. Devignes, D. C. Emily Koo, R. Bonneau, V. Gligorijević, M. Barot, H. Fang, S. Toppo, E. Lavezzo, M. Falda, M. Berselli, S. C. Tosatto, M. Carraro, D. Piovesan, H. U. Rehman, Q. Mao, S. Zhang, S. Vucetic, G. S. Black, D. Jo, D. J. Larsen, A. R. Omdahl, L. W. Sagers,

E. Suh, J. B. Dayton, L. J. McGuffin, D. A. Brackenridge, P. C. Babbitt, J. M. Yunes, P. Fontana, F. Zhang, S. Zhu, R. You, Z. Zhang, S. Dai, S. Yao, W. Tian, R. Cao, C. Chandler, M. Amezola, D. Johnson, J.-M. Chang, W.-H. Liao, Y.-W. Liu, S. Pascarelli, Y. Frank, R. Hoehndorf, M. Kulmanov, I. Boudellioua, G. Politano, S. Di Carlo, A. Benso, K. Hakala, F. Ginter, F. Mehryary, S. Kaewphan, J. Björne, H. Moen, M. E. E. Tolvanen, T. Salakoski, D. Kihara, A. Jain, T. Šmuc, A. Altenhoff, A. Ben-Hur, B. Rost, S. E. Brenner, C. A. Orengo, C. J. Jeffery, G. Bosco, D. A. Hogan, M. J. Martin, C. O'Donovan, S. D. Mooney, C. S. Greene, P. Radivojac and I. Friedberg, The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens, *bioRxiv* (2019).

5. R. You, Z. Zhang, S. Zhu, F. Sun, Y. Xiong and H. Mamitsuka, GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank, *Bioinformatics* **34**, 2465 (03 2018), [PubMed:29522145] [doi:10.1093/bioinformatics/bty130].
6. R. You, S. Yao, H. Mamitsuka and S. Zhu, DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction, *Bioinformatics* **37**, i262 (07 2021).
7. R. You, S. Yao, Y. Xiong, X. Huang, F. Sun, H. Mamitsuka and S. Zhu, NetGO: improving large-scale protein function prediction with massive network information, *Nucleic Acids Research* **47**, W379 (05 2019), [PubMed:31106361] [PubMed Central:PMC6602452] [doi:10.1093/nar/gkz388].
8. R. You, X. Huang and S. Zhu, Deeptext2go: Improving large-scale protein function prediction with deep semantic text representation, *Methods* **145**, 82 (2018), Data mining methods for analyzing biological data in terms of phenotypes.
9. V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho and R. Bonneau, Structure-based protein function prediction using graph convolutional networks, *Nature Communications* **12**, p. 3168 (May 2021).
10. B. Lai and J. Xu, Accurate protein function prediction via graph attention networks with predicted structure information, *Briefings in Bioinformatics* **23** (12 2021), bbab502.
11. M. Kulmanov and R. Hoehndorf, DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms, *Bioinformatics* **38**, i238 (06 2022).
12. M. Kulmanov, M. A. Khan and R. Hoehndorf, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, *Bioinformatics* **34**, 660 (10 2017), [PubMed:29028931] [PubMed Central:PMC5860606] [doi:10.1093/bioinformatics/btx624].
13. M. Kulmanov and R. Hoehndorf, DeepGOPlus: improved protein function prediction from sequence, *Bioinformatics* (07 2019), [PubMed:31350877] [doi:10.1093/bioinformatics/btz595].
14. Y. Cao and Y. Shen, TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding, *Bioinformatics* **37**, 2825 (03 2021).
15. T. Pan, C. Li, Y. Bi, Z. Wang, R. B. Gasser, A. W. Purcell, T. Akutsu, G. I. Webb, S. Imoto and J. Song, PFresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships, *Bioinformatics* **39** (02 2023), btad094.
16. Z. Wu, M. Guo, X. Jin, J. Chen and B. Liu, CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction, *Bioinformatics* (03 2023), btad123.
17. S. Yao, R. You, S. Wang, Y. Xiong, X. Huang and S. Zhu, NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information, *Nucleic Acids Research* **49**, W469 (05 2021).
18. C. Zhao, T. Liu and Z. Wang, Panda2: protein function prediction using graph neural networks, *NAR Genomics and Bioinformatics* **4** (January 2022).
19. W. Xiang, Z. Xiong, H. Chen, J. Xiong, W. Zhang, Z. Fu, M. Zheng, B. Liu and Q. Shi, Fapm: functional annotation of proteins using multimodal models beyond structural modeling, *Bioinformatics* **40** (November 2024).
20. G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin and B. Ghanem, Camel: Communicative

agents for ”mind” exploration of large language model society, in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

21. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur *et al.*, A large-scale evaluation of computational protein function prediction, *Nature methods* **10**, 221 (2013).
22. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. A. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaszner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjorne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. E. Sternberg, N. Skunca, F. Supek, M. Bosnjak, P. Panov, S. Dzeroski, T. Smuc, Y. A. I. Kourmpetis, A. D. J. van Dijk, C. J. F. t. Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney and I. Friedberg, A large-scale evaluation of computational protein function prediction, *Nat Meth* **10**, 221 (January 2013), [PubMed:23353650] [PubMed Central:PMC3584181] [doi:10.1038/nmeth.2340].
23. B. Buchfink, C. Xie and D. H. Huson, Fast and sensitive protein alignment using diamond, *Nature Methods* **12**, 59 EP (Nov 2014), [PubMed:25402007] [doi:10.1038/nmeth.3176].
24. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**, 1123 (2023).
25. Q. Yuan, J. Xie, J. Xie, H. Zhao and Y. Yang, Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion, *Briefings in Bioinformatics* **24**, p. bbad117 (03 2023).
26. S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V. Crichlow, K. Dalenberg, J. M. Duarte, S. Dutta, M. Fayazi, Z. Feng, J. W. Flatt, S. Ganesan, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. P. Hudson, I. Khokhriakov, C. L. Lawson, Y. Liang, R. Lowe, E. Peisach, I. Persikova, D. W. Piehl, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, B. Vallat, M. Voigt, B. Webb, J. D. Westbrook, S. Whetstone, J. Y. Young, A. Zalevsky and C. Zardecki, 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**, p. D488–D508 (November 2022).