SCIENCE CHINA
Information Sciences

• LETTER •

November 2026, Vol. 69, Iss. 11, 219102:1–219102:2
https://doi.org/10.1007/s11432-025-5060-9

A multimodal learning framework for protein function prediction based on multi-grained joint embedding

Jiangyi SHAO[1,2], Ziwen WANG[1], Zixu CHEN[1,2] & Bin LIU[1,2*]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[2]*Zhongguancun Academy, Beijing 100094, China*

Received 24 November 2025/Revised 17 March 2026/Accepted 25 April 2026/Published online 16 October 2026

Citation Shao J Y, Wang Z W, Chen Z X, et al. A multimodal learning framework for protein function prediction based on multi-grained joint embedding. Sci China Inf Sci, 2026, 69(11): 219102, https://doi.org/10.1007/s11432-025-5060-9

Protein function prediction is fundamental for understanding biological processes and elucidating cellular mechanisms [1, 2]. As proteins participate in nearly all life activities, such as metabolism, signal transduction, and molecular synthesis, accurate functional annotation is essential for studies in drug discovery and disease treatment [3]. However, experimental identification of protein functions is labor-intensive and time-consuming, making computational prediction an essential alternative.

With the rapid development of deep learning, a series of deep learning-based methods have been proposed to predict protein functions from large-scale biological data. Sequence-based models learn contextual representations directly from amino-acid sequences [4], while structure-based models leverage spatial information from predicted or experimental 3D structures [5]. Hybrid approaches combining these modalities, such as TransFun [6] and related graph-learning methods [7], have further enhanced prediction accuracy. However, most existing methods still rely on limited information sources and struggle to integrate heterogeneous biological knowledge, restricting their generalization.

We propose the multi-grained joint embedding and build a novel multimodal learning framework for protein function prediction, named MJEGO. The framework integrates sequence, structural, functional, and evolutionary information to learn coordinated representations across residue-, domain-, and protein-level hierarchies. By using an equivariant graph neural network and a cross-attention mechanism, the model fuses these modalities into a unified representation space. On the benchmark dataset, our method achieves consistent performance improvements over existing approaches. These results demonstrate that the proposed multi-grained joint embedding captures complementary biological information and achieves more accurate and robust functional predictions across all sub-ontology categories.

Methods. To overcome the limitations of single-modality approaches, we propose a multi-grained joint embedding model that hierarchically integrates fine-, medium-, and coarse-grained biological information. This framework captures complementary structural, functional, and evolutionary characteristics to construct a unified protein representation for function prediction. It combines four major information sources: protein sequence, three-dimensional structure, Gene Ontology (GO) annotation, and multiple sequence alignment (MSA) (Figure 1). Detailed information on this analysis can be found in Appendix A.8.

Results. To comprehensively evaluate the proposed multi-grained joint embedding multimodal learning framework, we conducted experiments on the benchmark dataset from TransFun, which contains proteins with a maximum sequence similarity of 30%.

The evaluation proceeds in three progressive stages: (i) overall performance comparison with state-of-the-art methods to validate the effectiveness of our framework; (ii) an ablation study on fine-grained representation to assess the impact of different sequence-structure encoders; and (iii) an ablation study on multi-grained joint embedding to examine how different information fusion strategies influence performance.

We compared our method with TransFun [6], GVP [5], and MSA Transformer [4] (Table C1). These models differ substantially in the types of biological information they exploit: TransFun integrates sequence and structure features, GVP relies solely on structural information, and MSA Transformer incorporates only MSA. This diversity of baselines enables a comprehensive evaluation of how well the proposed multimodal and multi-grained approach captures complementary biological cues. Detailed information on this analysis can be found in Appendix C.

The ablation study on fine-grained representation evaluates the impact of different backbone encoding strategies (Table C2). Our proposed framework consistently outperforms TransFun, revealing its robustness and adaptability to diverse feature representations. Among these, the Swin-Transformer achieves the most balanced and reliable results, particularly in CC and MF tasks, where structural features play a crucial role. Detailed information on this analysis can be found in Appendix C.

The ablation study on multi-grained fusion evaluates different fusion strategies (Table C3). The "GO first, then MSA" strategy yields the best overall performance on CC and MF, indicating that establishing a functional semantic foundation with GO information before incorporating evolutionary constraints leads to more coherent representations. Notably, the "Only GO" model achieves the best BP performance. This observation reflects a fundamental characteristic of the BP sub-ontology: BP annotations describe high-level cellular pathways and system-level func-

* Corresponding author (email: bliu@bliulab.net)

© Science China Press 2026

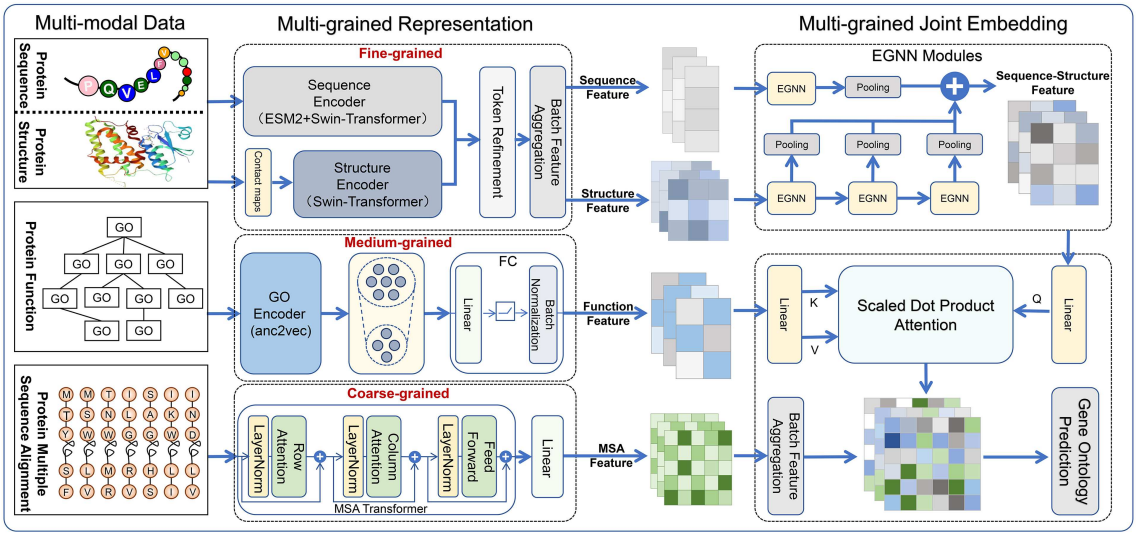

Figure 1 (Color online) Overall architecture of MJEGO. The framework integrates protein sequence, structure, GO semantic embeddings, and MSA across three granularity levels. Fine-grained: sequence embeddings (S-PLM) and AlphaFold2 coordinates are jointly encoded by EGNN. Medium-grained: GO-conditioned cross-modal attention fuses functional semantics with sequence-structure representations. Coarse-grained: MSA features are concatenated with the fused output for prediction via a two-layer MLP. All inputs are identical during training and inference; GO embeddings are ontology-derived priors, not ground-truth labels, introducing no information leakage.

tions that depend on hierarchical functional relationships across the GO graph. GO embeddings from Anc2Vec directly encode such ancestor-descendant dependencies, providing strong semantic signals well-suited to BP prediction. In contrast, MSA primarily captures residue-level evolutionary conservation, which is highly informative for CC and MF where local biochemical properties and structural context dominate, but may introduce noise for BP by emphasizing sequence-level conservation patterns that do not align with process-level functional abstractions. Detailed information on this analysis can be found in Appendix C.

These results reveal two key insights: (i) functional and evolutionary features contribute complementary perspectives, with GO providing functional semantics and MSA encoding evolutionary conservation; (ii) the fusion order significantly influences prediction performance, and a biologically motivated hierarchical strategy outperforms naive concatenation.

Discussion. In this study, we propose MJEGO, a multi-grained joint-embedding framework for protein function prediction that integrates sequence, structure, GO, and MSA information. Our method achieves competitive performance compared with existing models, including TransFun, GVP, and MSA Transformer.

This study makes two primary contributions. First, it introduces a multi-grained representation framework that explicitly models protein biology across three abstraction levels: residue (fine), protein (medium), and protein family (coarse). Unlike existing multimodal approaches that encode each modality independently and combine them through late fusion (e.g., concatenation), MJEGO employs a GO-conditioned cross-modal attention mechanism in which functional ontology actively guides the learning of sequence-structure representations. Second, it proposes a hierarchical fusion strategy with an empirically validated integration order. The ablation analysis demonstrates that the "GO first, then MSA" order consistently outperforms the reverse for CC and MF, revealing that establishing a functional semantic foundation before incorporating evolutionary constraints leads to more coherent representations. This order-sensitive hierarchical conditioning is fundamentally different from architecture-agnostic module stacking and constitutes a key methodological contribution beyond the individual components employed.

Future work may explore integrating more advanced geometric deep learning modules, such as 3D graph transformers or equivariant attention networks, to enhance spatial reasoning capability. In addition, extending the framework to incorporate dynamic or condition-specific omics data, such as expression or interaction profiles, could further refine functional inference and expand the model's applicability to broader biological and biomedical contexts.

Acknowledgements This work was supported by Beijing Zhongguancun Academy (Grant No. 02012502), National Natural Science Foundation of China (Grant No. 62502033), and China Postdoctoral Science Foundation (Grant No. 2024M764144).

Supporting information Appendixes A–C and Tables C1–C3. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

Availability of data and materials All source code, raw data, and scripts required to reproduce the results of this study are publicly available at the GitHub repository: https: //github.com/bliulab/MJEGO.

References

1 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
2 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630: 493–500
3 Bileschi M L, Belanger D, Bryant D H, et al. Using deep learning to annotate the protein universe. Nat Biotechnol, 2022, 40: 932–937
4 Rao R M, Liu J, Verkuil R, et al. MSA Transformer. In: Proceedings of International Conference on Machine Learning, 2021. 8844–8856
5 Jing B, Eismann S, Suriana P, et al. Learning from protein structure with geometric vector perceptrons. In: Proceedings of the International Conference on Learning Representations, 2021
6 Boadu F, Cao H, Cheng J. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. Bioinformatics, 2023, 39: i318–i325
7 Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins Struct Funct Bioinf, 2020, 88: 397–413