Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press



Balancing gene ontology annotation specificity in protein function prediction based on the protein sequence large graph

Jiangyi Shao, Shutao Chen, Ziwen Wang, et al.

Genome Res. published online April 15, 2026
Access the most recent version at doi:10.1101/gr.280816.125

P<P Published online April 15, 2026 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Balancing Gene Ontology annotation specificity in protein function prediction based on the protein sequence large graph

Jiangyi Shao[1,2], Shutao Chen[1], 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.

Correspondence and requests for materials should be addressed to B.L.

(Email: bliu@bliulab.net)

Abstract

Accurate protein function prediction is fundamental to advancing drug discovery, precision medicine, and understanding complex biological systems. While Gene Ontology (GO) provides a standardized framework for protein annotation, a critical challenge persists: the imbalance between low-specificity GO terms and high-specificity GO terms. This imbalance creates blind spots in our understanding of protein function landscapes, particularly in clinically relevant pathways. We present ProGO-PSL, a novel large graph architecture designed to resolve this imbalance. ProGO-PSL simultaneously leverages explicit domain identifier from InterPro and implicit evolutionary context from Multiple Sequence Alignments, fusing these complementary data sources within a powerful imbalance learning framework. Our model consistently outperforms state-of-the-art methods by 5-15% across all specificity levels and on both benchmark dataset and independent test set, demonstrating robust generalization. Furthermore, ProGO-PSL generates interpretable representations that clarify relationships between low- and high-specificity GO terms, enabling a more complete functional characterization of the proteome. This work accelerates the identification of therapeutic targets in previously uncharacterized biological pathways.

Keywords: protein function prediction, gene ontology annotation, protein sequence large graph, hybrid information fusion, imbalance learning

Introduction

The generation and functional expression of protein sequences is a complex process

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

involving the translation of nucleic acid sequences into proteins, which then fold into their functional forms (Bershtein et al. 2017). Protein Gene Ontology (GO) annotation plays a crucial role in bridging genetic information to protein function, enabling the construction of fine-grained mappings between proteins and their associated GO terms (Ashburner et al. 2000; Nei et al. 2010; Peng et al. 2020; Kulenkampff et al. 2021). Specifically, Gene Ontology (GO) contains a standardized vocabulary (term) for describing the biological roles, molecular functions, and cellular localizations of genes and their products (proteins). These annotations are organized into three main ontology categories: Cellular Component (CC), which specifies the locations within a cell where proteins perform their functions; Molecular Function (MF), which describes the biochemical activities of proteins; and Biological Process (BP), which represents larger processes composed of molecular activities (Harris et al. 2004; Bepler and Berger 2021; Gene Ontology et al. 2023). Despite the availability of large databases such as UniProt (Bateman et al. 2023), which provide annotations for tens of millions of proteins, these annotated entries represent only a limited portion of the sequence diversity captured in resources such as UniRef (Suzek et al. 2007; Suzek et al. 2015). This gap highlights a fundamental limitation in current approaches: the mapping from protein sequences to functional annotations remains incompletely characterized, particularly in the presence of large numbers of unannotated sequences (Vu et al. 2023).

A primary challenge in protein function prediction is the imbalance in annotation specificity. This issue arises from the hierarchical structure of the Gene Ontology (GO), a Directed Acyclic Graph (DAG) in which terms range from low-specificity to high-specificity (Ashburner et al. 2000). GO terms near the root of the hierarchy represent low-specificity, high-frequency annotations, whereas GO terms deeper in the graph describe precise, specialized functions that are sparsely annotated across the proteome. Consequently, prediction models are often biased toward the easily learned, low-specificity terms, leading to poor performance on the rarer yet often more informative specific functions (Durairaj et al. 2023). Overcoming this bias is critical, as specific annotations are essential for uncovering novel biological pathways and identifying precise targets for drug development (Fox et al. 2023; Pavlopoulos et al. 2023; Reggiardo et al. 2023). Therefore, this study focuses on achieving balanced prediction performance across the GO hierarchy, (Soding 2017; Michaelis et al. 2023; Alfatah et al. 2024; Nomburg et al. 2024), ensuring that these functionally critical annotations are not overlooked. We refer to this phenomenon as “specificity imbalance,” encompassing both annotation density imbalance and information content asymmetry across GO terms.

This gap is characterized by an imbalanced distribution of information, with well-studied proteins richly annotated with low-specificity GO terms, and sparsely studied proteins associated with specialized, less frequent terms. A series of computational methods has been proposed to improve the protein-GO mapping to bridge this gap. DeepGOPlus (Kulmanov and Hoehndorf 2020) combines multi-scale- features using one-dimensional convolutional neural networks to enhance the functional information embedded in protein sequences. However, relying solely on the information from a

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

single sequence is insufficient for predicting GO annotations at a fine granularity. TALE (Cao and Shen 2021) on the other hand, uses a pre-trained language model to increase the amount of useful information for GO annotations. Similarly, DeepFRI (Gligorijevic et al. 2021) combines representations derived from pretrained sequence models with structural information, allowing functional signals to be captured at residue-level resolution. Recent studies have emphasized the importance of association data in improving GO predictions. DeepGraphGO (You et al. 2021) leverages protein-protein interaction (PPI) data to predict GO terms, as functional interactions often imply relationships at the GO level. PFresGO (Pan et al. 2023) combines a protein language model with the hierarchical structure of GO terms, enhancing the functional information extracted from sequences and improving the accuracy of some specific GO term predictions.

DeepGO-SE(Kulmanov et al. 2024), DeepGOZero (Kulmanov and Hoehndorf 2022), DPFunc(Wang et al. 2025) represent three recent approaches that have advanced protein function prediction under different modeling assumptions. DeepGO-SE [31] frames the task as approximate semantic entailment, generating multiple approximate models of GO, and aggregating predictions across these models to exploit GO's axiomatic structure. Recent state-of-the-art methods have increasingly incorporated InterPro domains (Paysan-Lafosse et al. 2023) as explicit features. These domains are detected by the InterProScan tool (Jones et al. 2014), which analyzes protein sequences to identify conserved domains. Using the resulting domain identifiers as categorical features provides more discriminative signals for distinguishing functional similarities between different sequences. For examples, DeepGOZero (Kulmanov and Hoehndorf 2022) incorporates ELEmbeddings (Kulmanov et al. 2019) with InterPro domain identifiers to learn associations between sequence-derived domain identifiers and GO terms from training data. DPFunc (Wang et al. 2025) also incorporates InterPro domain identifiers to guide attention mechanisms in its architecture so as to focus on structurally and functionally important regions of proteins.

Despite these advances, existing methods share a common limitation: they do not explicitly account for the imbalance in annotation density across GO specificity levels. In the GO hierarchy, terms at lower specificity levels (closer to the root) tend to accumulate substantially more annotations than terms at higher specificity levels (deeper in the hierarchy), because general functions are easier to assign experimentally and are inherited by all child terms through the true path rule. As a result, models trained on such data are biased toward predicting low-specificity terms, where training signal is abundant. However, these models underperform on high-specificity terms that carry more precise and biologically informative annotations. Moreover, most sequence-based methods implicitly assume that sequence similarity is a reliable indicator of functional similarity. This assumption becomes less reliable at higher levels of specificity, where subtle functional distinctions may not be necessarily reflected in overall sequence similarity. Consequently, this specificity imbalance reflects a structural challenge that cannot be resolved solely by improving sequence representations or incorporating

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

additional features (Beltran et al. 2025).

Motivated by these observations, we develop a protein sequence large graph (PSL) framework that improves prediction across GO specificity levels through a dual-pathway design. Specifically, it combines implicit evolutionary contexts derived from multiple sequence alignments (MSA) with domain-level constraints obtained from InterPro identifiers. While existing methods such as DeepGOZero and PFresGO incorporate domain annotations or ontology structure, they typically rely on a single representation source or treat proteins uniformly across different specificity levels. In contrast, ProGO-PSL learns explicit and implicit representations through separate pathways and integrates them via adaptive gating, enabling the model to balance precision and generalization according to the specificity of the target GO terms. Within an imbalance-aware learning framework, this design allows ProGO-PSL to capture functional relationships across both low- and high-specificity GO terms while maintaining robustness across diverse protein sequences.

Results

Characterization of the specificity imbalance

To empirically characterize the specificity imbalance in GO annotation, we analyzed the protein-function landscape across all three ontology categories. As illustrated in Fig. 1A, proteins annotated with low-specificity GO terms form densely connected clusters, whereas proteins annotated with high-specificity terms are sparsely represented: a pattern consistent across CC, MF, and BP. This uneven annotation density results in an information asymmetry: low-specificity terms, though broadly assigned, carry limited discriminative information, while high-specificity terms encode precise functional distinctions but are underrepresented in training data and thus harder to model.

A separate but compounding factor concerns the relationship between sequence similarity and functional similarity. Analysis of representative protein pairs with specific annotations (Fig. 1B) showed that high sequence similarity frequently fails to predict comparable functional similarity, particularly at higher GO specificity levels (Fig. 1C). Quantification of information content further confirmed this divergence (Fig. 1D), indicating a systematic mismatch between sequence-space proximity and function-space proximity.

These observations suggest three contributing factors: (i) annotation imbalance across GO terms, where specific functions are systematically underrepresented relative to general ones; (ii) study bias across proteins, where well-characterized organisms dominate existing annotations; and (iii) representation mismatch between sequence space and function space, where sequence proximity does not reliably indicate functional proximity. Collectively, these observations suggest that the core challenge of GO annotation is not merely class imbalance. Rather, the task requires learning a mapping from protein sequence space to a hierarchically structured function space

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

under conditions where high-specificity annotations are both sparse and informationally distinct from low-specificity ones. This motivates a task formulation that explicitly accounts for this hierarchical specificity structure..

Problem formulation

The preceding analysis identified three factors underlying the specificity imbalance: annotation sparsity at high specificity levels, study bias across proteins, and a systematic mismatch between sequence-space proximity and function-space proximity. We now formalize the annotation task in a way that makes these challenges explicit.

We define Gene Ontology (GO) annotation as a hierarchical multi-label classification problem (Kabir and Shehu 2022). Let $\mathbf{x} = [A_1, A_2, \ldots, A_L]$ denote a protein sequence of length L, where each A_i represents one of the 21 standard amino acids(Yang et al. 2020). Let $\mathcal{G} = (V, E)$ denote the directed acyclic graph (DAG) of a given GO category (CC, MF, or BP), where $V = \{t_1, t_2, \ldots, t_m\}$ is the set of GO terms and E encodes the hierarchical "is-a" relationships between terms. The task is to learn a mapping $f: \mathbf{x} \to \mathbf{y} \in \{0,1\}^m$, where $y_j = 1$ indicates that the protein is annotated with term t_j (Pan et al. 2023). This mapping must satisfy the true-path constraint: if $y_j = 1$ and t_k is an ancestor of t_j in $\mathcal{G}$, then $y_k = 1$.

The specificity imbalance characterized above manifests in this formulation as two interrelated difficulties. First, the marginal label frequency $P(y_j = 1)$ decreases sharply with the depth of t_j in $\mathcal{G}$, creating a long-tailed distribution over the output space that biases standard loss functions toward low-specificity terms. Second, the mapping f cannot be reliably approximated by sequence similarity alone, because proximity in sequence space does not entail proximity in the function space implied by $\mathcal{G}$, particularly at higher specificity levels where functional distinctions are fine-grained (Boadu et al. 2023). An effective model must therefore learn representations that capture functional relevance beyond sequence similarity and maintain predictive accuracy across different specificity levels of the ontology hierarchy (Sapoval et al. 2022).

Model architecture

Information content analyze

To quantify the informativeness of Gene Ontology (GO) terms and guide the development of our computational model, we employed the Information Content (IC) metric (Lin 1998b; Schlicker et al. 2006; Gligorijevic et al. 2021; Kulmanov and Hoehndorf 2022; Unsal et al. 2022) (**Fig. 2**). As shown in **Fig. 2A**, the IC distribution of GO terms follows a long-tailed pattern, where most GO terms have high IC values, indicating that they are highly specific and contain rich information. However, a small subset of frequently occurring terms has lower IC values, reflecting their broad

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

applicability across various biological contexts. This long-tail distribution highlights an information imbalance: the majority of GO terms, which are highly specific and informative, represent specialized functions, while a small number of low-specificity, broadly applicable terms represent general functions.

To better understand how this long-tail distribution impacts computational modeling, we further investigated its underlying structure and the challenges it presents. Low-specificity GO terms are frequently annotated across diverse proteins because they appear in a wide range of biological contexts, making them less-specific in terms of distinguishing unique protein functions (i.e., lower IC values). In contrast, high-specificity GO terms (with high IC values), are annotated less frequently, making it difficult for predictive models to accurately capture them. Especially, due to the inherent data imbalance that skews model performance toward the more frequently occurring low-specificity terms. This challenge underscores a significant disparity in information content between protein sequences and their corresponding GO terms, particularly across different specificity levels. As a result, a more nuanced computational approach is required to effectively bridge this gap and improve the accuracy of GO term prediction across all specificity levels.

To address the imbalance across GO specificity levels, we construct the protein sequence large graph (PSL), which employs an imbalance learning framework to integrate both explicit and implicit information. We compare the performance of different types of informative data (explicit, implicit, and hybrid) for predicting GO terms as shown in **Fig. 2B**. Explicit information, derived from curated resources, provides discriminative signals that are particularly effective for resolving fine-grained functional distinctions. In contrast, implicit information, derived from multiple sequence alignments, that captures subtle evolutionary patterns that captures broad evolutionary patterns that support generalization across diverse functional contexts. As shown in **Fig. 2B**, our results confirm this hypothesis: models using only a single information source exhibit biases, whereas the hybrid model fusing both explicit and implicit information achieves superior and more balanced performance. These effects are not exclusive to specific regions of the ontology but instead reflect differences in how each information source contributes to prediction under varying levels of functional granularity. This fusion is therefore crucial for accurately predicting terms across the full spectrum of specificity. In particular, it helps mitigate the trade-off between performance on low-specificity terms, which tend to be more extensively annotated, and high-specificity terms, which are comparatively sparse in available annotations.

In summary, the information content metric reveals an intrinsic information asymmetry between proteins and GO terms across different specificity levels. To address this, we propose a hybrid information-based approach that constructs a protein sequence large graph based on hybrid information.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Protein sequence large graph

By integrating both explicit and implicit information, we aim to use an effective imbalance learning framework to predict Gene Ontology (GO) terms from protein sequences. This approach will help overcome the long-tail distribution challenge, ensuring accurate prediction of both high- and low-IC GO terms. To achieve this, we propose a Protein Sequence Large graph (PSL) model, called ProGO-PSL, which consists of three key components: a deep residual channel attention network (RCAN) (Jie Hu 2018; Wang et al. 2020; Ridnik et al. 2021), a constrained embedding representation (CER) (Kulmanov and Hoehndorf 2022), and a fusion of representations (FoR) (Shazeer 2020).

As shown in **Fig. 2C**, the RCAN captures implicit information by extracting features from multiple sequence alignments, generating representations that reflect the broader functional landscape of proteins. The CER is a multi-layer perceptron that learns from InterPro domain identifiers (Kulmanov and Hoehndorf 2022; Paysan-Lafosse et al. 2023) to capture explicit features, producing representations that are closely aligned with known functional categories. The FoR is a Transformer-based architecture that uses Gated Linear Units (GLUs) (Shazeer 2020) to fuse the implicit representations from RCAN with the explicit representations from CER. This architecture balances the prediction performance of low-specificity and high-specificity, which is crucial for hierarchical multi-label classification.

Datasets and Evaluation methods

Benchmark dataset. The benchmark dataset was constructed based on the UniProt/SwissProt (released in April 2022), which is a manually curated resource widely used for protein function prediction. Following the protocol employed by DeepGOPlus (Kulmanov and Hoehndorf 2020), we constructed a high-confidence dataset from UniProt/SwissProt (April 2022 release) by retaining only proteins with experimentally supported annotations (e.g., EXP, IDA, IPI). The resulting dataset contains 79,230 proteins and is aligned with the Gene Ontology release of October 2022 (31,253 reviewed terms).

To ensure a consistent evaluation protocol (Kulmanov and Hoehndorf 2020), the dataset was partitioned into training, validation, and test subsets (85.5%, 9.5%, and 5%, respectively). Model development, including parameter tuning and selection, was performed using the training and validation sets, while final performance was assessed on the held-out test set.

Independent test set. To further evaluate generalization under realistic conditions, we constructed an independent test set consisting of proteins added to UniProt/SwissProt between May 2022 and March 2025. These proteins do not overlap with the benchmark dataset and therefore provide a strict assessment of model transferability to newly annotated sequences.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Our model trained with the training set of the benchmark dataset was directly performed on the independent test set to give the prediction results.

The statistical information of both the benchmark dataset and the independent test set was listed in the **Supplemental Table S1**.

Performance comparison across different specificity level

The information content (Lin 1998b; Schlicker et al. 2006; Gligorijevic et al. 2021; Kulmanov and Hoehndorf 2022; Unsal et al. 2022) of a Gene Ontology (GO) term is positively correlated with its depth in the hierarchical structure of the ontology, where terms deeper in the hierarchy tend to have higher IC values. Therefore, we evaluated the prediction accuracy of GO terms at different levels of information content. We compared ProGO-PSL against several state-of-the-art methods, including DeepGO-SE(Kulmanov et al. 2024), DeepGOZero(Kulmanov and Hoehndorf 2022), ESM-2 (Lin et al. 2023), as well as other competitive methods for GO annotation (The detailed configurations are provided in **Supplemental Table S3**).

To evaluate performance, we grouped the GO terms based on their information content into two categories: IC $\geq$ 10, representing highly specific terms, and IC $<$ 10, covering less specific terms. The evaluation metrics used were Fmax (maximum F-score) and AuPRC (Area under the Precision-Recall Curve), and the results are provided for Cellular Component (CC), Molecular Function (MF), and Biological Process (BP) ontologies. **Fig. 3** summarizes the results for each method, comparing performance for the two IC groups as well as for all GO terms combined.

Performance on benchmark dataset

For the benchmark dataset, we evaluated performance using a global hold-out strategy, holding out 5% of proteins for testing. **Fig. 3A** shows results across two information content (IC) groups (IC $\geq$ 10 for high-specificity terms, IC $<$ 10 for low-specificity terms) as well as all GO terms combined.

ProGO-PSL showed consistent performance across GO categories relative to existing approaches (Buchfink et al. 2015; Steinegger and Soding 2017; Kulmanov and Hoehndorf 2022; Lin et al. 2023; Kulmanov et al. 2024; van Kempen et al. 2024). For highly specific terms (IC $\geq$ 10), the some methods also showed strong results. DeepGO-SE achieved an Fmax of 0.738 and DeepGOZero 0.736 on the Cellular Component (CC) ontology. Both methods benefit from encoding GO axioms into geometric embeddings, which aligns well with the relatively well-structured cellular localization hierarchy. However, ProGO-PSL achieved an Fmax of 0.543 on the Biological Process (BP) ontology, surpassing DeepGO-SE (0.501). This because BP terms describe multi-step processes involving diverse proteins and pathways. DPFunc, although designed to highlight functional regions within protein structures through domain-guided attention, performed poorly (0.261 in BP). These results are consistent with the hypothesis that

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

detecting functional regions in structures alone is insufficient for predicting complex biological processes, which inherently involve multiple proteins and their interactions.

For low-specificity GO terms (IC < 10), the results highlight differences in how various approaches integrate functional information. DeepGOZero performed well in BP (Fmax = 0.544), reflecting its learning capacity, which propagates predictions across definition axioms connecting broad and specific functions. DeepGO-SE showed reduced performance in the same setting (Fmax = 0.512), suggesting that approximate semantic entailment is less effective for coarse-grained categories. DPFunc again underperformed (Fmax = 0.279), indicating limited ability to generalize beyond specific structural motifs. ProGO-PSL achieved the best BP performance in this group (Fmax = 0.553), showing that its large graph-based framework can capture both narrow and broad functional associations by integrating diverse forms of protein similarity.

When evaluating all GO terms together, ProGO-PSL demonstrated clear advantages. In Molecular Function (MF), ProGO-PSL reached an Fmax of 0.717, compared with 0.646 for DeepGO-SE and 0.683 for DeepGOZero. The difference was even more striking in BP, where ProGO-PSL achieved 0.560 compared with 0.484 for DeepGO-SE and 0.517 for DeepGOZero. These results highlight that they may fail to capture additional biological relationships. The imbalance learning on large graph allows ProGO-PSL to combine explicit domain knowledge with implicit evolutionary information, bridging this gap. The underperformance of DPFunc (Fmax = 0.344 in MF) reinforces the idea that while domain-guided attention enhances interpretability, it lacks the breadth required for general prediction across the heterogeneous mechanisms of protein function.

Performance on independent test set

To further evaluate the generalization ability of ProGO-PSL and avoid data leakage, various methods were evaluated on an independent test set constructed based on UniProt/SwissProt. All the proteins in this independent test set were added into UniProt/SwissProt during May 2022 and March 2025, which is fully independent from the benchmark data set constructed based on UniProt/SwissProt released on April 2022. The results of different methods are shown in **Fig. 3B**.

This validation exposes fundamental differences in robustness. DeepGO-SE showed marked performance decline when moving to this kind of data. Its MF performance on all terms fell from 0.646 to 0.580. Approximate semantic entailment relies on multiple "world models" of GO axioms generated during training. Once GO annotations expand with new experimental data, these models become outdated, limiting generalization. DeepGOZero experienced even greater declines, particularly in MF (0.683 to 0.611). While its zero-shot mechanism is advantageous for rare terms during training, it becomes a liability when new proteins introduce functions not easily explained by pre-existing definition axioms. DPFunc performed poorly across all ontologies (e.g., BP all terms Fmax = 0.244), suggesting that domain-guided structural features are insufficient to handle newly characterized proteins.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

By contrast, ProGO-PSL maintained both accuracy and stability. For highly specific terms, it achieved the highest Fmax scores across all ontologies (0.721 in CC, 0.690 in MF, 0.517 in BP). For low-specificity terms, its performance remained strong and showed minimal decline compared with the benchmark dataset (0.739 in CC, 0.672 in MF, 0.527 in BP). The margin of improvement over other methods widened under this evaluation. In MF (all terms), ProGO-PSL reached 0.736, compared with 0.580 for DeepGO-SE and 0.611 for DeepGOZero, corresponding to gains of 27% and 20%. In BP (all terms), ProGO-PSL achieved 0.542, compared with 0.428 for DeepGO-SE and 0.466 for DeepGOZero. These results indicate that our large graph model effectively integrate both implicit and explicit information, adapts better to proteins discovered in future.

In conclusion, these results highlight a key computational challenge in protein function prediction: methods that embed fixed biological knowledge can perform well, but their strengths are conditional. Language-model-based approaches, such as DeepGO-SE and ESM-2, exploit pretrained sequence embeddings to capture broad functional information. In contrast, domain-based approaches, such as DeepGOZero and DPFunc, incorporate InterPro domain identifiers as explicit features to improve learning of sequence–function associations. However, their performance drops sharply on newly added proteins. ProGO-PSL encodes knowledge in the large-graph structure of protein sequences rather than relying on fixed axiomatic or domain-specific rules. This design enables the model to propagate functional information across diverse relationships, uncover indirect associations, and generalize to novel sequences. Such adaptability is particularly critical for the BP ontology, where biological processes represent emergent network-level properties rather than attributes of individual proteins.

ProGO-PSL is a Robust and Specific Protein Function Predictor

A central challenge in protein function prediction is the long-tailed distribution of Gene Ontology (GO) annotations, where most terms are highly specific and sparsely annotated. ProGO-PSL addresses this bottleneck by integrating hybrid information from protein sequences and large graphs. To evaluate its ability to generalize across ontologies and maintain specificity in difficult cases, we assessed performance on the independent test set using three complementary perspectives: i) specific grouped annotation. We split the GO terms into two groups based on information content ($IC \geq 10$ and $IC < 10$) and assessed performance for each group individually and combined; ii) hierarchical grouped annotation. We grouped the GO terms by both their depth in the GO hierarchy and their information content; and iii) fine-grained grouped annotation (see **Supplemental Table S2** for detailed statistics). We categorized GO terms into numerical IC intervals and evaluated performance across each interval (Unsal et al. 2022). The results are summarized in **Fig. 4**.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

In the specific grouped annotation (**Fig. 4A**), we benchmarked ProGO-PSL against four state-of-the-art methods (DeepGO-SE, DeepGOZero, DPFunc, and ESM2) across three annotation specificity levels. ProGO-PSL demonstrated superior precision-recall trade-offs, achieving the highest Fmax scores in both molecular function (MF) and cellular component (CC) ontologies. For high-specificity terms (IC ≥ 10), ProGO-PSL attained Fmax values of 72.09% (CC) and 69.04% (MF), substantially outperforming the next-best methods DeepGO-SE (70.81% in CC) and DeepGOZero (66.70% in MF). In the biological process (BP) ontology, characterized by greater annotation heterogeneity and functional complexity, ProGO-PSL maintained competitive accuracy (Fmax = 51.67%). Notably, for low-specificity terms (IC < 10), where annotation density enables most methods to achieve high recall, ProGO-PSL sustained exceptional performance with Fmax values of 83.32% (CC), 94.90% (MF), and 66.07% (BP), demonstrating robust generalization across the specificity spectrum.

The results of hierarchical grouped annotation (**Fig. 4B**) revealed ProGO-PSL's consistent superiority across six distinct GO term categories defined by the intersection of hierarchical depth (shallow/medium/deep) and information content (high/low). ProGO-PSL's performance advantage was particularly pronounced for deep, high-IC terms, where annotation is sparse, and methods typically degrade. For example, in BP, ProGO-PSL was the best-performing method across all six subgroups, achieving Fmax scores of 92.72% (shallow/low), 73.26% (shallow/high), 54.07% (medium/low), 56.34% (medium/high), 53.04% (deep/low), and 51.74% (deep/high). This uniform dominance contrasts with ESM-2 and DeepGO-SE, which showed relative strength in shallow or low-IC subsets but deteriorated substantially in deep/high-IC categories. ProGO-PSL's ability to maintain balanced performance in deep/high-IC BP terms—where annotation is sparse and complexity greatest—demonstrates its capacity to capture long-range functional dependencies in the GO hierarchy.

In the fine-grained grouped annotation (**Fig. 4C**), Matthews Correlation Coefficient (MCC) analysis and the threshold analysis (Chicco and Jurman 2020) across numerical IC intervals elucidated complementary strengths among methods while confirming ProGO-PSL's overall superiority. In MCC analysis, for BP terms, ESM-2 performed best only in the lowest IC interval (IC < 5; MCC = 0.93 vs. ProGO-PSL 0.92). Beyond this range, ProGO-PSL consistently achieved the highest MCC, with values rising from 0.33 at IC 5–6 to 0.61 at IC 6–7, exceeding both DeepGO-SE (0.27–0.41) and DeepGOZero (0.29–0.42). Performance remained stable across higher IC bins, with ProGO-PSL achieving 0.54 MCC for IC ≥ 15. In MF, ProGO-PSL achieved near-perfect discrimination at IC < 5 (MCC = 0.997) and sustained superiority through IC 8–9 (0.784 vs. 0.310 for DeepGO-SE and 0.280 for DeepGOZero). because GO term annotation is determined by applying an optimal threshold that maximizes Fmax, the distribution of thresholds directly reflects the discriminative power of a model. In BP, ProGO-PSL's thresholds were both narrowly distributed and stable across IC bins (0.68–0.82, median 0.73), indicating well-calibrated decision boundaries that reliably separate positives from negatives across specificity levels.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

These results confirm that ProGO-PSL achieves the state-of-the-art performance in both robustness and specificity. Language-model-based approaches, such as DeepGO-SE and ESM-2, leverage pretrained embeddings to capture broad functional patterns, but their performance declines for highly specific terms. Domain-based approaches, such as DeepGOZero and DPFunc, incorporate InterPro domain identifiers to learn sequence-function associations effectively, yet their reliance on predefined domain features limits generalization. In contrast, ProGO-PSL learns functional associations directly from sequence data by its large graph framework. This design enables ProGO-PSL to maintain balanced performance across both low-specificity and high-specificity GO terms. Such balanced performance is particularly critical for the BP ontology, where annotations often reflect emergent biological processes at the network level rather than localized attributes of individual proteins.

Effective Integration of Complementary Hybrid Information in Protein Sequence Large Graph

The hybrid information employed by the Protein Sequence Large Graph (PSL) leverages the complementary strengths of explicit and implicit information. This complementarity arises from the distinct yet synergistic roles that each type of information plays in improving Gene Ontology (GO) term prediction. To demonstrate this, we conducted experiments analyzing the contributions of explicit and implicit information separately and in combination. We evaluated the model's ability to discriminate between positive and negative samples and plotted the distribution of these prediction results. **Fig. 5** visually illustrates the complementary nature of these two types of information.

In **Fig. 5A**, we compare the prediction performance of the hybrid information model with models that use only explicit or implicit information. Using Average Precision (AP) scores across different GO terms as the evaluation metric, we observe that the hybrid model's performance closely mirrors that of the explicit information model. This similarity suggests that explicit information contributes to improved discriminative performance, as reflected by its ability to reduce false positives, primarily by improving the detection of negative samples. This higher precision is crucial for reducing false positives, thereby optimizing the model's classification performance. The fact that the hybrid model preserves this precision advantage indicates that the fusion mechanism retains the discriminative contribution of explicit information, rather than diluting it through indiscriminate combination with implicit features.

In **Fig. 5B**, we further examine the hybrid model's discriminative ability by comparing the score distributions for positive and negative samples. The score distribution of the implicit information model closely aligns with that of the hybrid model, indicating that implicit information plays a critical role in improving the model's

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

generalization capability. Specifically, implicit information helps the model better identify positive samples by capturing subtle, abstract relationships that are not immediately apparent from explicit data alone. This contributes to the model's robustness and adaptability across diverse contexts, enhancing its overall predictive power. Similarly, the alignment between the hybrid and implicit models in score distribution demonstrates that the fusion process preserves the generalization strength of implicit information. Taken together, Figs. 5A and 5B show that the hybrid model simultaneously preserves the precision of explicit information and the generalization capacity of implicit information. This complementary effect cannot be achieved by either information source alone and reflects the effectiveness of the adaptive fusion strategy.

In summary, the integration of explicit and implicit information in the Protein Sequence Large Graph proves to be highly effective. This can be summarized in three key points: i) explicit information primarily enhances precision by improving the detection of negative samples, reducing false positives; ii) implicit information enhances generalization by better identifying positive samples, improving the model's robustness; and iii) the fusion of these two types of information through a GLU-based Transformer enables the model to combine their complementary strengths. Compared with simple concatenation or fixed-weight fusion, this adaptive integration allows the model to balance precision and generalization across GO terms with different specificity levels.

Feature analysis

This study suggests that the specificity level imbalance arises from the long-tailed distribution of information content in Gene Ontology (GO) terms. To address this disparity, ProGO-PSL employs an imbalance learning framework to integrate explicit and implicit information based on Protein Sequence Large Graph (PSL). One of the key advantages of this approach is its ability to generate interpretable representations, allowing for a deeper investigation into the relationships between highly specific and low-specificity GO terms. Through feature analysis, we demonstrated the impact of the specificity level imbalance on GO annotation and showcased how ProGO-PSL effectively addresses this issue, as shown in **Fig. 6**.

Our comprehensive model analysis reveals how ProGO-PSL addresses the specificity level imbalance in protein function prediction. Using t-SNE visualization of model weights (Fig. 5A), we demonstrate that while DeepGOPlus struggles with high-specificity GO terms due to their long-tailed distribution, ProGO-PSL effectively captures the global relationships between GO terms across different information content levels. To evaluate performance with varying sequence similarities, we employed Spearman correlation coefficients to measure both specificity in predicting high-specificity GO terms (Spearman-1) and generalizability across all terms

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

(Spearman-2). The results (Fig. 5B) confirm that high-specificity GO terms remain rare regardless of sequence similarity, creating prediction biases in conventional methods including ESM2. ProGO-PSL overcomes this limitation by effectively combining explicit and implicit information, enabling more accurate annotation across diverse proteins. Furthermore, by comparing protein similarity calculated by our model against similarity derived from actual GO annotations (Fig. 5C), we observed a positive linear correlation that validates ProGO-PSL's ability to reduce the imbalance in protein-GO association information. This three-part analysis collectively demonstrates how our approach successfully bridges the gap between protein sequences and their GO terms, providing a more balanced representation of the protein function landscape.

In **Fig. 6A**, the weights of the final fully connected layers of DeepGOPlus (Kulmanov and Hoehndorf 2020) and ProGO-PSL were analyzed using the t-SNE algorithm (van der Maaten and Hinton 2008). These weights represent the distribution of learned GO terms within the respective models. We highlighted GO terms with high information content, located deep in the GO hierarchy (referred to as high-specificity GOs), for easier identification. The figure shows that DeepGOPlus (Kulmanov and Hoehndorf 2020) struggles to capture the relationships between high-specificity GO terms and low-specificity terms. This is largely due to the long-tailed distribution of information content, which traditional methods fail to address. In contrast, ProGO-PSL effectively captures the global relationships between GO terms at different levels of information content, enabling a more comprehensive understanding of the distributional properties across different GO specificity levels.

In **Fig. 6B**, we divided protein sequences based on different sequence similarities, as proteins with varying sequence similarities exhibit different levels of information content. We then evaluated and compared the performance of ESM2 (Lin et al. 2023), Foldseek (van Kempen et al. 2024), and other computational methods in predicting GO terms for proteins with different sequence similarities. We quantified the agreement between predicted and reference annotations using Spearman's correlation coefficient (Spearman 2010), which captures monotonic relationships between the two distributions. The Spearman-1 score denotes the performance in predicting high-specificity GO terms, while the Spearman-2 score reflects the performance across all GO terms. These scores provide insight into the specificity (Spearman-1) and generalizability (Spearman-2) of the models. The results indicate that high-specificity GO terms are rare, regardless of sequence similarity, leading to prediction biases in most computational methods, including ESM2. In contrast, ProGO-PSL successfully combines explicit and implicit information, enabling more accurate GO annotation for proteins with varying sequence similarities and information content levels.

In **Fig. 6C**, we compared protein similarity calculated by ProGO-PSL with protein similarity derived from GO annotations. Protein similarity based on GO annotation represents true functional similarity, while the similarity calculated by ProGO-PSL represents predicted functional similarity. We used the Spearman score (ρ) to measure the correlation between the predicted and true similarity distributions. The positive

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

linear correlation observed between the two distributions indicates that ProGO-PSL effectively reduces the imbalance in protein-GO association information. This demonstrates the model's ability to bridge the gap between protein sequences and their corresponding GO annotations.

Discussion

Our study provides a refined understanding of the specificity imbalance in Gene Ontology (GO) annotation. Rather than being independent or isolated, high-specificity GO terms are structurally and functionally connected to low-specificity terms within the GO hierarchy. This observation suggests that functional specificity emerges from hierarchical dependencies across annotation levels, rather than from local sequence similarity alone. Consequently, approaches that rely solely on sequence similarity or treat GO terms as independent labels are inherently limited in capturing the full structure of protein function space. These findings reframe the core challenge of GO annotation as learning a mapping between protein sequence space and a hierarchically structured function space, under conditions where proximity in sequence space does not necessarily correspond to proximity in function space.

Building on this understanding, ProGO-PSL explicitly models the mismatch between protein sequence space and the hierarchically structured function space defined by GO. The model integrates two complementary sources of information: implicit evolutionary signals from multiple sequence alignments and explicit domain-level constraints from InterPro identifiers. This enables the model to capture both global functional relationships and local structural signals, which is particularly important for high-specificity GO terms that are underrepresented in training data. A potential limitation of MSA-derived features is that evolutionary signals may reinforce misleading homology when sequence similarity does not reliably reflect functional similarity. To mitigate this effect, ProGO-PSL learns refined representations from reconstructed alignment data and conditions them on complementary domain-level features. This design reduces over-reliance on homology alone. Importantly, the model does not rely on predefined mappings between InterPro annotations and GO terms; instead, domain features are treated as sequence-derived inputs, and their associations with function are learned directly from data. This design ensures that the model performs function prediction rather than simply redistributing existing annotations across the GO hierarchy.

The interpretability of the learned representations further supports this mechanism. The alignment between predicted protein similarities and GO-derived functional similarities suggests that ProGO-PSL partially bridges sequence-based representations and functional organization. This bridging reflects the complementary roles of explicit and implicit information. Explicit features primarily contribute discriminative precision, which is particularly important when resolving fine-grained functional distinctions often associated with high-specificity terms. In contrast, implicit features support

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

generalization across broader regions of the functional space, contributing to stable performance on lower-specificity terms and the overall distribution. More broadly, future advances may depend less on improving single modalities and more on frameworks that model the relational structure of biological function.

Despite these improvements, several limitations remain. First, the model depends on the availability and quality of multiple sequence alignments and InterPro annotations. For proteins with limited homologous sequences or incomplete domain annotations, the quality of both implicit and explicit representations may be reduced. Second, although ProGO-PSL improves predictions for high-specificity GO terms, extremely rare terms with minimal annotation data remain challenging due to the inherent sparsity of the training distribution. Third, the current framework does not explicitly enforce the true-path constraint of the GO hierarchy during training or inference; incorporating ontology-aware structural constraints could further improve prediction consistency across parent-child term pairs. Finally, the evaluation benchmarks, while augmented by the independent test set, remain predominantly composed of well-characterized proteins from model organisms, and generalizability to underrepresented taxa warrants systematic investigation.

These findings have broader implications for computational protein function prediction. The observed mismatch between sequence similarity and functional similarity highlights the need to move beyond sequence-only approaches toward models that explicitly account for hierarchical and contextual relationships in biological systems. In this context, integrating additional modalities, such as protein structural information from resources like AlphaFold and contextual embeddings from protein language models, may further enhance the representation of functional space. More broadly, this study demonstrates that explicitly modeling the hierarchical structure of the annotation space, rather than treating it as a flat classification problem, can improve predictive performance. This observation suggests that similar strategies may be beneficial for other biological ontologies, such as disease ontologies, pathway databases, and phenotype classification systems, which also exhibit long-tailed distributions and hierarchical dependencies. Taken together, these results point toward a general direction in which protein function prediction may benefit from the integration of multi-level biological evidence within ontology-aware learning frameworks, providing a more balanced and biologically consistent representation of functional space.

Methods

Information content

We quantify GO term specificity using Shannon information content (IC), defined as $\mathrm{IC}(x) = -\log(|x|/N)$, where $|x|$ denotes the number of proteins annotated with term x and N is the total number of proteins in the dataset. Under this formulation, terms

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

associated with fewer proteins correspond to higher specificity. Details are provided in Supplemental Material S1.3.

Multiple sequence alignment encoding

MSA profiles capture implicit evolutionary information but may contain noise from non-homologous matches. To extract robust representations, we first convert the MSA into continuous embeddings by mapping each amino acid and gap symbol to a learnable d-dimensional vector (Eq. S1). The embedded MSA is then processed through a series of convolutional operations, including down-sampling, residual blocks, and up-sampling, to reconstruct a denoised representation that captures residue frequency, co-occurrence, and positional patterns. Full architectural details and hyperparameters are provided in Supplemental Material S1.4.

Implicit information representation via Deep Residual Channel Attention Network

Inspired by trRosetta's use of deep convolution to project MSA data into structural feature space, we employ a Deep Residual Channel Attention Network (RCAN) (Marcel and Rodriguez 2010; He et al. 2016; Huang et al. 2023) to learn ontology-specific representations from the reconstructed MSA. The network consists of an initial convolutional down-sampling layer followed by four residual stages, each incorporating Efficient Channel Attention (ECA) modules (Wang et al. 2020). The ECA mechanism performs local cross-channel interaction via adaptive 1D convolution, allowing the network to weight channels by their informativeness without dimensionality reduction. The output is aggregated through global average pooling to produce a fixed-dimensional implicit representation r_{conv} for each protein. Architectural details and equations are provided in Supplemental Material S1.5.

Explicit information representation via constrained embedding representation

We derive explicit functional features from InterPro domain annotations (Paysan-Lafosse et al. 2023), obtained by scanning protein sequences with InterProScan (Jones et al. 2014; Paysan-Lafosse et al. 2023). Each protein is represented as a binary vector $\mathbf{v} \in \{0,1\}^N$, where $N$is the total number of distinct InterPro annotations and each entry indicates the presence or absence of a specific domain, family, or functional site annotation (Kulmanov and Hoehndorf 2022). Importantly, only the presence or absence of each InterPro entry is used as a categorical input feature; no textual descriptions, functional summaries, or pre-existing InterPro-to-GO term mappings are incorporated into the model. All associations between domain features and GO terms are learned entirely from the training data. This design ensures that the explicit pathway functions

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

as a learned feature extractor rather than an annotation translation system. This vector is projected to a fixed-dimensional space via a linear layer and reshaped into $g$groups of dimension d to form constrained embeddings: $c_c = \text{reshape}(vW_c + b_c; g, d)$ where each group captures a subset of domain-level constraints on protein function. Architectural details and equations are provided in Supplemental Material S1.6.

Fusion of representations

The implicit representation $r_{conv}$and the constrained embedding c_ca are fused using a Transformer architecture with a SwiGLU-based feedforward network. The implicit representation is first reshaped and linearly projected to match the shape of c_c, yielding r_p. The two embeddings are concatenated along the group dimension to form $x_c = \text{concat}(c_c, r_p) \in \mathbb{R}^{2g \times d}$.

This fused embedding is processed by a self-attention encoder in which the standard MLP-based feedforward network is replaced by a SwiGLU feedforward network (Ramachandran et al. 2018) :

$$\text{SwiGLU}(x) = \text{Swish}_\beta(x\mathbf{W} + b) \otimes (x\mathbf{V} + c) \quad (1)$$

$$\text{FFN}_{\text{SwiGLU}}(x) = (\text{Swish}_1(x\mathbf{W}) \otimes x\mathbf{V})\mathbf{W}_2 \quad (2)$$

$$\text{SelfAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{FFN}_{\text{SwiGLU}}\left(\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}\right) \quad (3)$$

where Equation 1 represents the original SwiGLU, x denotes the input, W and V denote the weights, b and c denote the bias, and β is the parameter of the Swish activation function. Equation 2 denotes the SwiGLU based feedforward network FFN_{SwiGLU} instead of MLP based feedforward network in this section, where $Swish_1$ denotes $\beta = 1$ in Swish activation function, and $\mathbf{W}$, $\mathbf{V}$, and $\mathbf{W}_2$ denote the weights. SelfAttn denotes the self-attention encoder based on SwiGLU, where the attention weights are computed from the similarity between projected feature groups in the fused representation, scaled by $\sqrt{d_k}$ for numerical stability, and normalized via the softmax operation. This formulation allows interactions between explicit and implicit feature groups within x_c, enabling the model to capture cross-source dependencies during fusion.

The SwiGLU gating mechanism enables the model to adaptively select which components of the fused representation are informative, providing a learned balance between explicit domain-level signals and implicit evolutionary signals. This is critical for mitigating the specificity imbalance, as the relative informativeness of explicit versus implicit features varies across specificity levels.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Imbalance learning

The specificity imbalance identified in the preceding analysis, where high-specificity GO terms are severely underrepresented relative to low-specificity terms. It poses a direct challenge to model training: standard training procedures tend to be dominated by the abundant low-specificity terms, resulting in poor predictive performance for the sparse but informationally rich high-specificity terms. To address this, we design a training framework comprising three complementary components: a two-step training strategy, an asymmetric loss function, and training stabilization techniques.

Three separate hierarchical multi-label classifiers are trained for CC, MF, and BP. Each classifier is trained in two sequential steps. In the first step, the model is trained using only the implicit representations derived from MSA via the RCAN pathway. Because evolutionary signals captured by MSA are broadly distributed across specificity levels, this step provides a robust initialization that is not dominated by any particular specificity range. In the second step, the explicit representations from InterPro domain annotations are incorporated, and the full model (implicit and explicit pathways with GLU-based Transformer fusion) is fine-tuned. Domain annotations provide direct functional constraints that are particularly useful for resolving fine-grained functional distinctions, which are more prevalent among high-specificity terms.

To further mitigate the label imbalance at the loss level, we employ the Asymmetric Loss (ASL) function (Ridnik et al. 2021) during both training steps:

$$\mathcal{L}_{ASL} = \begin{cases} (1-p)^{\gamma_+}\log(p) & \text{for positive GO terms } (GO^+) \\ (p_m)^{\gamma_-}\log(1-p_m) & \text{for negative GO terms } (GO^-) \end{cases} \quad (4)$$

where p denotes the predicted probability, and p_m introduces a margin-based truncation that suppresses low-confidence negative responses. The parameters γ_+ and γ_- control the relative emphasis on positive and negative samples, respectively, allowing the loss to focus on hard examples, which are particularly important for high-specificity GO terms that are sparsely represented. The margin m introduces a hard threshold that effectively discards easy negative samples (i.e., GO terms clearly absent for a given protein), while the asymmetric focusing parameters allow independent control over the down-weighting of easy positives versus easy negatives. This is critical for specificity imbalance: for high-specificity terms, the vast majority of training samples are negative, and without asymmetric treatment, the loss signal would be overwhelmed by these uninformative negatives. We set $\gamma_+ = \gamma_- = 4$and $m = 10^{-8}$.

Two techniques are employed to stabilize the training process, motivated by the concern that imbalanced label distributions may cause the model to converge toward solutions biased toward low-specificity terms.

First, we apply Exponential Moving Average (EMA) (Grill et al. 2020) to maintain a smoothed copy of the model parameters, with a decay rate of $\lambda = 0.9997$. EMA maintains a smoothed trajectory of model parameters, which is particularly relevant when training with imbalanced labels: without smoothing, gradient updates are disproportionately influenced by the frequent low-specificity terms that dominate each

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

mini-batch, and the resulting parameter oscillations can destabilize representations learned for rare high-specificity terms.

Second, we adopt the One Cycle Learning Rate (OneCycleLR) schedule (Smith and Topin 2018) with a peak learning rate of 2×10^{-4}. During the warm-up phase, the low learning rate allows the model to first establish stable representations of the more prevalent low-specificity terms. As the learning rate increases toward its peak, the model gains sufficient gradient magnitude to refine decision boundaries for the harder, less frequent high-specificity terms. The subsequent cosine annealing phase progressively reduces the learning rate to consolidate learned representations across all specificity levels.

All three ProGO-PSL models (CC, MF, BP) are trained independently with the Adam optimizer using default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$). Hyperparameters are tuned on the corresponding validation sets.

Datasets

No new data was generated in this study; all data is derived from publicly available resources. The experimental data used in this work is available at online prediction webserver of ProGO-PSL (http://bliulab.net/ProGO-PSL/). All data used in this work is from public resources the UniProt website and the Gene Ontology website. The benchmark dataset can be downloaded from http://bliulab.net/ProGO-PSL/download/swissprot, including the benchmark data, and the trained model weight. The independent test set can be downloaded from http://bliulab.net/ProGO-PSL/download/swissprot_ind. The Multiple Sequence Alignment (MSA) data of benchmark dataset can be downloaded from this link http://bliulab.net/ProGO-PSL/download/msas. The MSA data of independent test set can be downloaded from this link http://bliulab.net/ProGO-PSL/download/ind_msas.

Software availability

The ProGO-PSL framework is publicly available on GitHub (https://github.com/altriasjy31/pslg-server). The source code is also provided as Supplemental Code associated with this manuscript. Detailed documentation and usage instructions are available at http://bliulab.net/ProGO-PSL/document.

Acknowledgements

We would like to thank the three reviewers, whose insightful comments have obviously improved the presentation of this paper. This work was supported by the National Natural Science Foundation of China (No. 62325202, U22A2039 and 62502033), the China Postdoctoral Science Foundation (No. 2024M764144) and the

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Beijing Zhongguancun Academy (No. 20240310 and 20240101).

Author contributions

J.S. and B.L. conceived the idea. J.S. developed the method. B.L. and J.S. supervised the project. J.S. designed the experiments, performed the analyses and wrote the paper. Z. W helps to finish the experiment. B.L., S.C., Z.C., and J.S. provided critical feedback during the study and helped revise the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Reference

Alfatah M, Lim JJJ, Zhang YZ, Naaz A, Cheng TYN, Yogasundaram S, Faidzinn NA, Lin JJ, Eisenhaber B, Eisenhaber F. 2024. Uncharacterized yeast gene, an effector of TORC1 signaling in a mitochondrial feedback loop, accelerates cellular aging via - and -dependent mechanisms. *Elife* **12**.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25-29.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* **3**: 361-362.

Bateman A Martin MJ Orchard S Magrane M Ahmad S Alpi E Bowler-Barnett EH Britto R Cukura A Denny P et al. 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**: D523-D531.

Beltran A, Jiang Xe, Shen Y, Lehner B. 2025. Site-saturation mutagenesis of 500 human protein domains. *Nature* doi:10.1038/s41586-024-08370-4.

Bepler T, Berger B. 2021. Learning the protein language: Evolution, structure, and function. *Cell Systems* **12**: 654-669 e653.

Bershtein S, Serohijos AWR, Shakhnovich EI. 2017. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Curr Opin Struc Biol* **42**: 31-40.

Boadu F, Cao H, Cheng J. 2023. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Bioinformatics **39**: i318-i325.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**: 59-60.

Cao Y, Shen Y. 2021. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics* **37**: 2825-2833.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**: 6.

Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A et al. 2023. Uncovering new families and folds in the natural protein universe. *Nature* **622**: 646–653.

Fox BA, Urba WJ, Jensen SM, Page DB, Curti BD, Sanborn RE, Leidner RS. 2023. Cancer's Dark Matter: Lighting the Abyss Unveils Universe of New Therapies. *Clin Cancer Res* **29**: 2173-2175.

Gene Ontology C Aleksander SA Balhoff J Carbon S Cherry JM Drabkin HJ Ebert D Feuermann M Gaudet P Harris NL et al. 2023. The Gene Ontology knowledgebase in 2023. *Genetics* **224**.

Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H et al. 2021. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**.

Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M et al. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol 33 (ed. H Larochelle, et al.), pp. 21271-21284. Curran Associates, Inc.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258-D261.

He K, Zhang X, Ren S, Sun J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, doi:10.1109/CVPR.2016.90, pp. 770-778.

Huang Z, Chen S, Yu L. 2023. Predicting new drug indications based on double variational autoencoders. *Computers in Biology and Medicine* **164**: 107261.

Jie Hu LS, Gang Sun. 2018. Squeeze-and-Excitation Networks. In *CVPR*.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Kabir A, Shehu A. 2022. GOProFormer: A Multi-Modal Transformer Method for Gene Ontology Protein Function Prediction. *Biomolecules* **12**: 1709-1728.

Kulenkampff K, Perez AMW, Sormanni P, Habchi J, Vendruscolo M. 2021. Quantifying misfolded protein oligomers as drug targets and biomarkers in Alzheimer and Parkinson diseases. *Nat Rev Chem* **5**: 277-294.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Kulmanov M, Guzmán-Vega FJ, Duek Roggli P, Lane L, Arold ST, Hoehndorf R. 2024. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence* **6**: 220-228.

Kulmanov M, Hoehndorf R. 2020. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**: 422-429.

Kulmanov M, Hoehndorf R. 2022. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics* **38**: i238-i245.

Kulmanov M, Liu-Wei W, Yan Y, Hoehndorf R. 2019. EL Embeddings: Geometric Construction of Models for the Description Logic EL++. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, doi:10.24963/ijcai.2019/845, pp. 6103-6109. International Joint Conferences on Artificial Intelligence Organization.

Lin D. 1998a. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296-304. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Lin D. 1998b. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann Publishers Inc.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123-1130.

Marcel S, Rodriguez Y. 2010. Torchvision the Machine-Vision Package of Torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, doi:10.1145/1873951.1874254, pp. 1485-1488. Association for Computing Machinery, New York, NY, USA.

Michaelis AC, Brunner AD, Zwiebel M, Meier F, Strauss MT, Bludau I, Mann M. 2023. The social and structural architecture of the yeast protein interactome. *Nature* **624**: 192-200.

Nei M, Suzuki Y, Nozawa M. 2010. The Neutral Theory of Molecular Evolution in the Genomic Era. *Annu Rev Genom Hum G* **11**: 265-289.

Nomburg J, Doherty EE, Price N, Bellieny-Rabelo D, Zhu YK, Doudna JA. 2024. Birth of protein folds and functions in the virome. *Nature* doi:10.1038/s41586-024-07809-y: 710–717.

Pan T, Li C, Bi Y, Wang Z, Gasser RB, Purcell AW, Akutsu T, Webb GI, Imoto S, Song J. 2023. PFresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. *Bioinformatics* **39**.

Pavlopoulos GA, Baltoumas FA, Liu SR, Selvitopi O, Camargo AP, Nayfach S, Azad A, Roux S, Call L, Ivanova NN et al. 2023. Unraveling the functional dark matter through global metagenomics. *Nature* **622**: 594–602.

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

ML, Bork P, Bridge A, Colwell L et al. 2023. InterPro in 2022. *Nucleic Acids Res* **51**: D418-D427.

Peng C, Trojanowski JQ, Lee VMY. 2020. Protein transmission in neurodegenerative disease. *Nat Rev Neurol* **16**: 199-212.

Ramachandran P, Zoph B, Le QV. 2018. Searching for Activation Functions. In *6th International Conference on Learning Representations*, pp. 1-13, Vancouver Convention Center, Vancouver, BC, Canada.

Reggiardo RE, Maroli SV, Peddu V, Davidson AE, Hill A, LaMontagne E, Aaraj YA, Jain M, Chan SY, Kim DH. 2023. Profiling of repetitive RNA sequences in the blood plasma of patients with cancer. *Nature Biomedical Engineering* **7**: 1627-1635.

Ridnik T, Lawen H, Noy A, Ben E, Sharir BG, Friedman I. 2021. TResNet: High Performance GPU-Dedicated Architecture. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, doi:10.1109/WACV48630.2021.00144, pp. 1399-1408.

Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, Barberan CJ, Dannenfelser R, Dun C, Edrisi M et al. 2022. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* **13**: 1728.

Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**: 302.

Shazeer N. 2020. GLU Variants Improve Transformer. *arXiv* doi:https://arxiv.org/abs/2002.05202.

Smith LN, Topin N. 2018. Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates. In *6th International Conference on Learning Representations*, pp. 1-16, Vancouver Convention Center, Vancouver, BC, Canada.

Soding J. 2017. Big-data approaches to protein structure prediction. *Science* **355**: 248-249.

Spearman C. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology* **39**: 1137-1150.

Steinegger M, Soding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**: 1026-1028.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282-1288.

Suzek BE, Wang YQ, Huang HZ, McGarvey PB, Wu CH, Consortium U. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**: 926-932.

Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T. 2022. Learning functional properties of proteins with language models. *Nature Machine Intelligence* **4**: 227-245.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**: 2579-2605.

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. 2024. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**: 243-246.

Vu MH, Akbar R, Robert PA, Swiatczak B, Sandve GK, Greiff V, Haug DTT. 2023. Linguistically inspired roadmap for building biologically reliable protein language models. *Nature Machine Intelligence* **5**: 485-496.

Wang JZ, Du ZD, Payattakool R, Yu PS, Chen CF. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**: 1274-1281.

Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, doi:10.1109/CVPR42600.2020.01155, pp. 11531-11539.

Wang W, Shuai Y, Zeng M, Fan W, Li M. 2025. DPFunc: accurately predicting protein function via deep learning with domain-guided structure information. *Nature Communications* **16**: 70.

Yang JY, Anishchenko I, Park H, Peng ZL, Ovchinnikov S, Baker D. 2020. Improved protein structure prediction using predicted interresidue orientations. *P Natl Acad Sci USA* **117**: 1496-1503.

You RH, Yao SW, Mamitsuka H, Zhu SF. 2021. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics* **37**: I262-I271.

Figure 1. Overview of the specificity imbalance problem in gene ontology annotation. This figure illustrates how protein function prediction is challenged by disparities across Gene Ontology (GO) specificity levels and by the disconnect between sequence similarity and functional similarity. (A) Force Atlas 2 layout (Bastian et al. 2009) of protein-protein and GO-GO networks (is a relationship), where highly connected nodes are colored red, others are colored blue. Four representative protein pairs are selected for detailed analysis. (B) We select four representative proteins and their GO terms. (C) Comparison of sequence similarity (computed with MMseqs2 (Steinegger and Soding 2017)) and functional similarity (calculated with Wang's algorithm (Wang et al. 2007)) for the representative protein pairs, highlighting their divergence. (D) Scatter plots quantifying imbalance: the first three show GO term distribution by information content (IC) and hierarchical depth, while the last one compares sequence similarity with GO term similarity (using Lin's algorithm (Lin 1998a; Unsal et al. 2022)

Figure 2. Overview of our hybrid information-based approach via the protein sequence large graph. We illustrated our protein sequence large graph model through the lens of information content. (A) We used the information content metric to

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

analyze the imbalanced distribution across GO terms of different specificity levels. The first two parts of panel A visualize the distribution of information content in Gene Ontology (GO) terms and proteins, respectively, where the orange bars represent high-specificity elements, and the blue bars represent low-specificity ones. The third part of panel A shows the distribution of information content in a hierarchical GO graph. The fourth part of panel A depicts the information gap between proteins and GO terms based on their similarity, where protein sequence similarity is calculated using MMseqs2 (Steinegger and Soding 2017) and GO similarity is based on the hybrid information-based metric (Lin's approach) from this study (Lin 1998a; Unsal et al. 2022). (B) We illustrated the characteristics of the different computational models based on their corresponding input types and compared their performance in distinguishing negative and positive samples. (C) We presented the pipeline of our protein sequence large graph based on hybrid information

Figure 3. Performance comparison across different specificity levels. We evaluate ProGO-PSL against state-of-the-art computational methods for predicting GO terms with varying specificity levels across Cellular Component (CC), Molecular Function (MF), and Biological Process (BP) ontologies. GO terms are grouped by information content: IC $\geq$ 10 (highly specific terms) and IC $<$ 10 (low specific terms). (A) Performance on benchmark dataset. (B) Performance on independent test set.

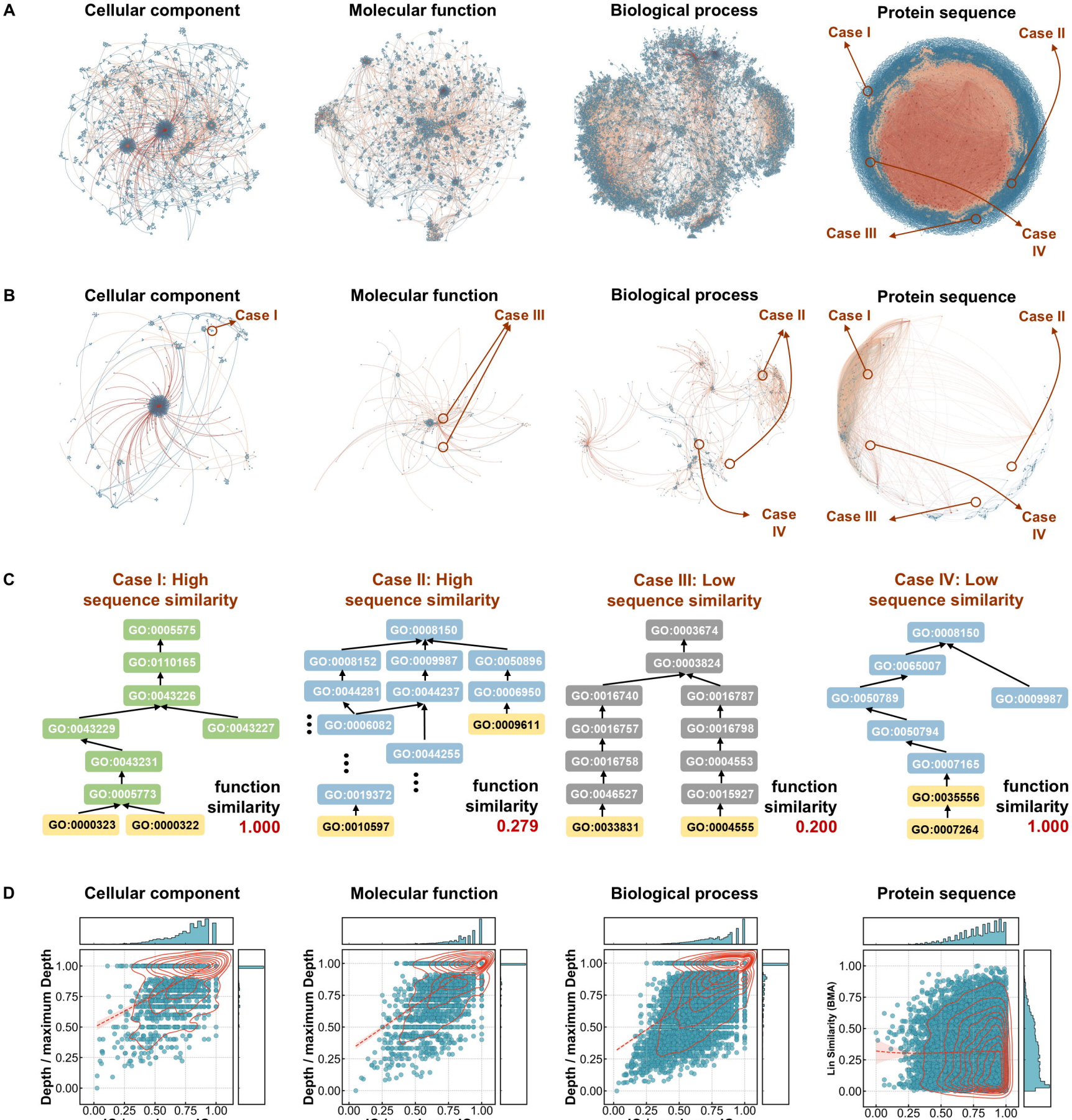
Figure 4. Performance evaluation of robustness and specificity on independent test set. We evaluated and displayed the prediction performance of ProGO-PSL on independent test set via three experiments. The prediction performance of ProGO-PSL is compared with the other state-of-the-art methods through three experiments on an independent test set. (A) Precision-Recall (PR) curves with Iso-F1 contours (Gligorijevic et al. 2021) are used to compare ProGO-PSL with other computational methods. (B) Radar plots display the performance of ProGO-PSL against other methods for six distinct groups of GO terms, which are categorized by a combination of information content (high/low) and hierarchical depth (shallow/medium/deep). (C) The performance across fine-grained IC intervals is analyzed in two parts. The upper panel shows the Matthews Correlation Coefficient (MCC) scores for each method as line plots, along with the size of each IC interval represented by bar plots. The lower panel illustrates the distribution of predicted scores for positive and negative samples using violin plots, overlaid with the optimal F1-score threshold curve for ProGO-PSL.

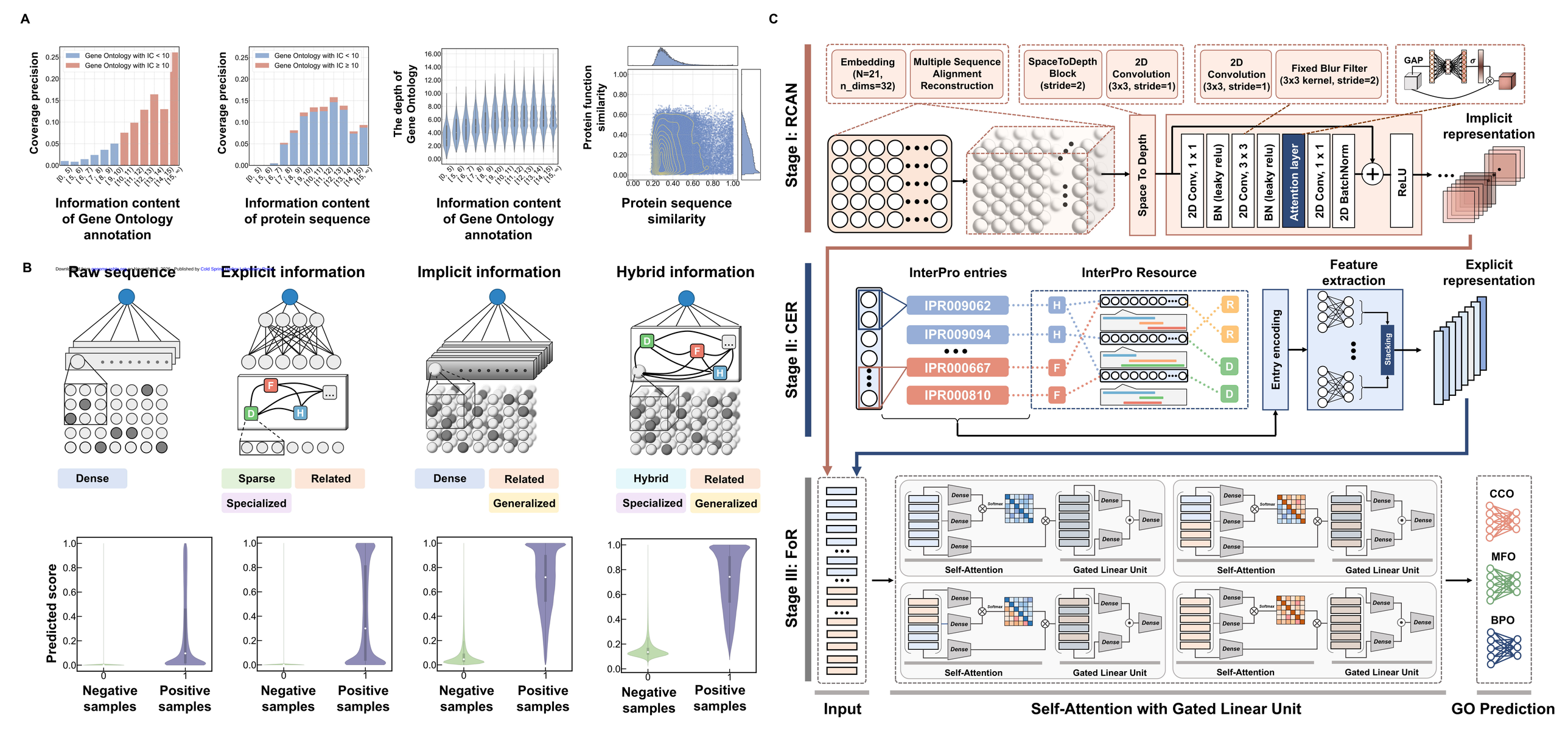
Figure 5. Accurate prediction of gene ontology annotation using the complementary property of hybrid information. We compare the prediction results of different types of information to reveal the complementary property. (A) We plot a scatter diagram where each point denotes an average precision score of a Gene Ontology term and the coordinate of point is the prediction score pair. We also calculated the p-value of the compared two results by the Wilcoxon ranked method.

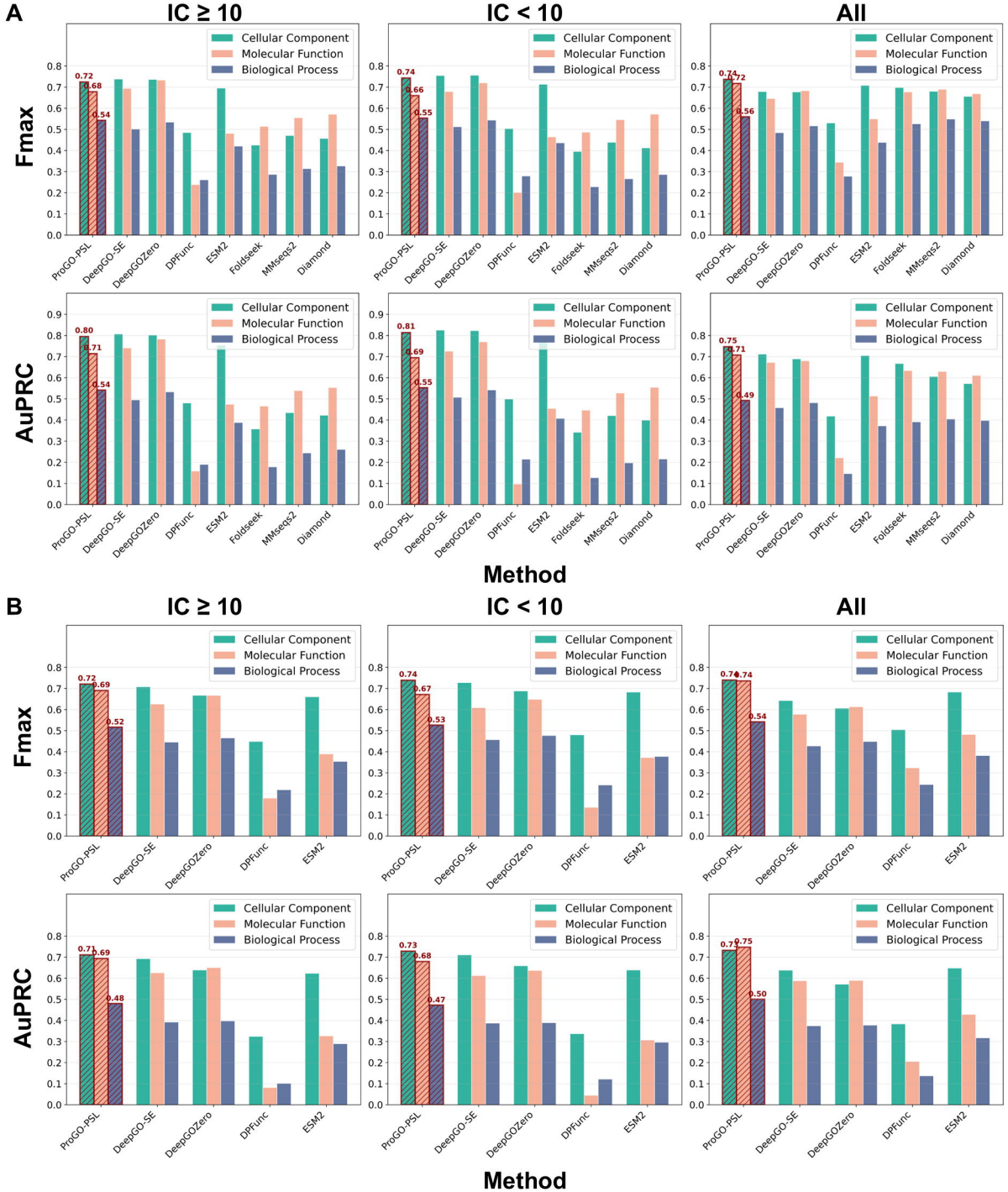
Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

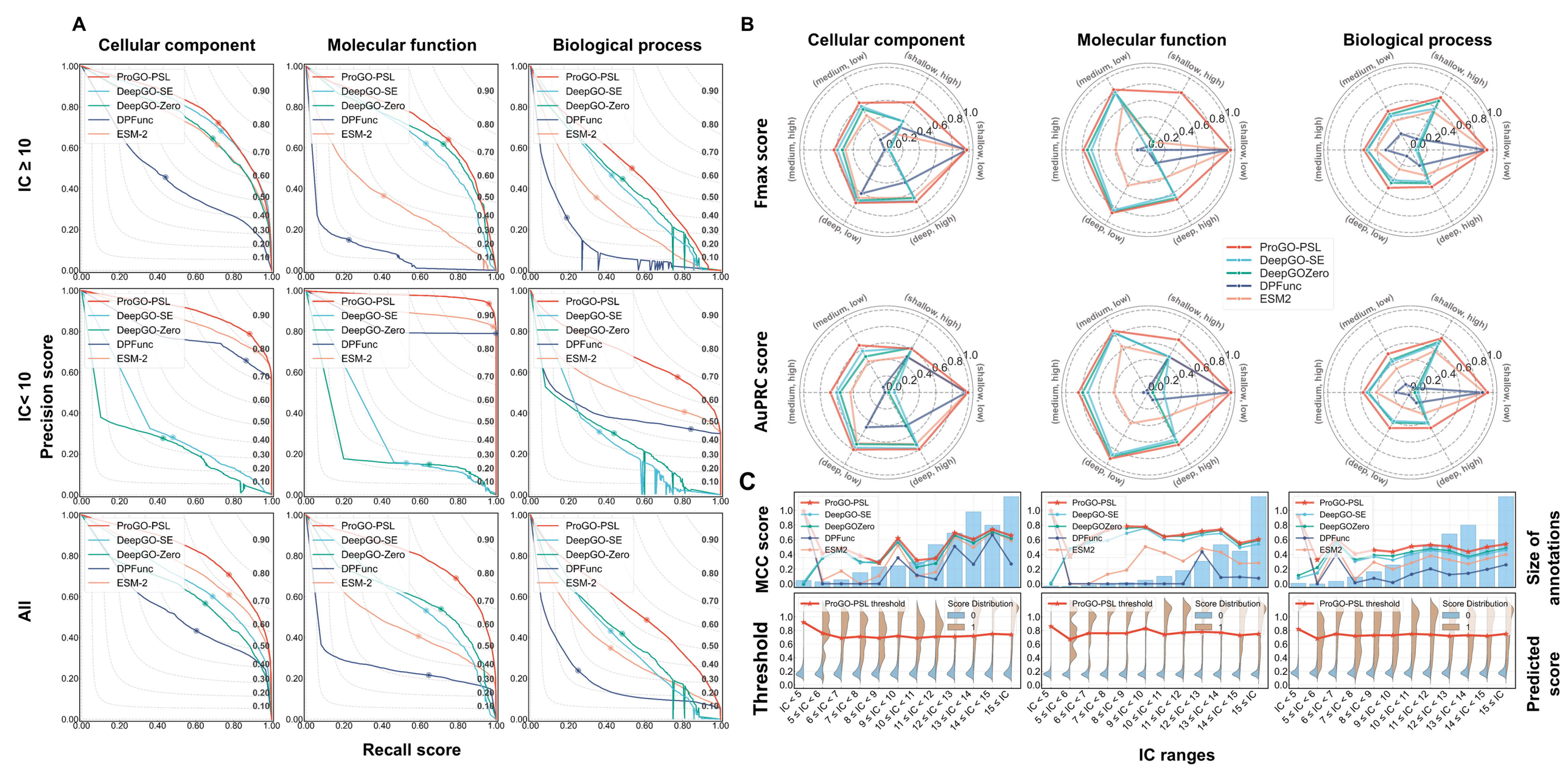
(B) We displayed the prediction results of distinguishing negative and positive samples. The first two parts denote the results before the two types of information are feed forward into the protein sequence large graph. The third part denotes the results of integrating the two types of information initially. The last part denotes the final prediction results of the GLU-based Transformer.

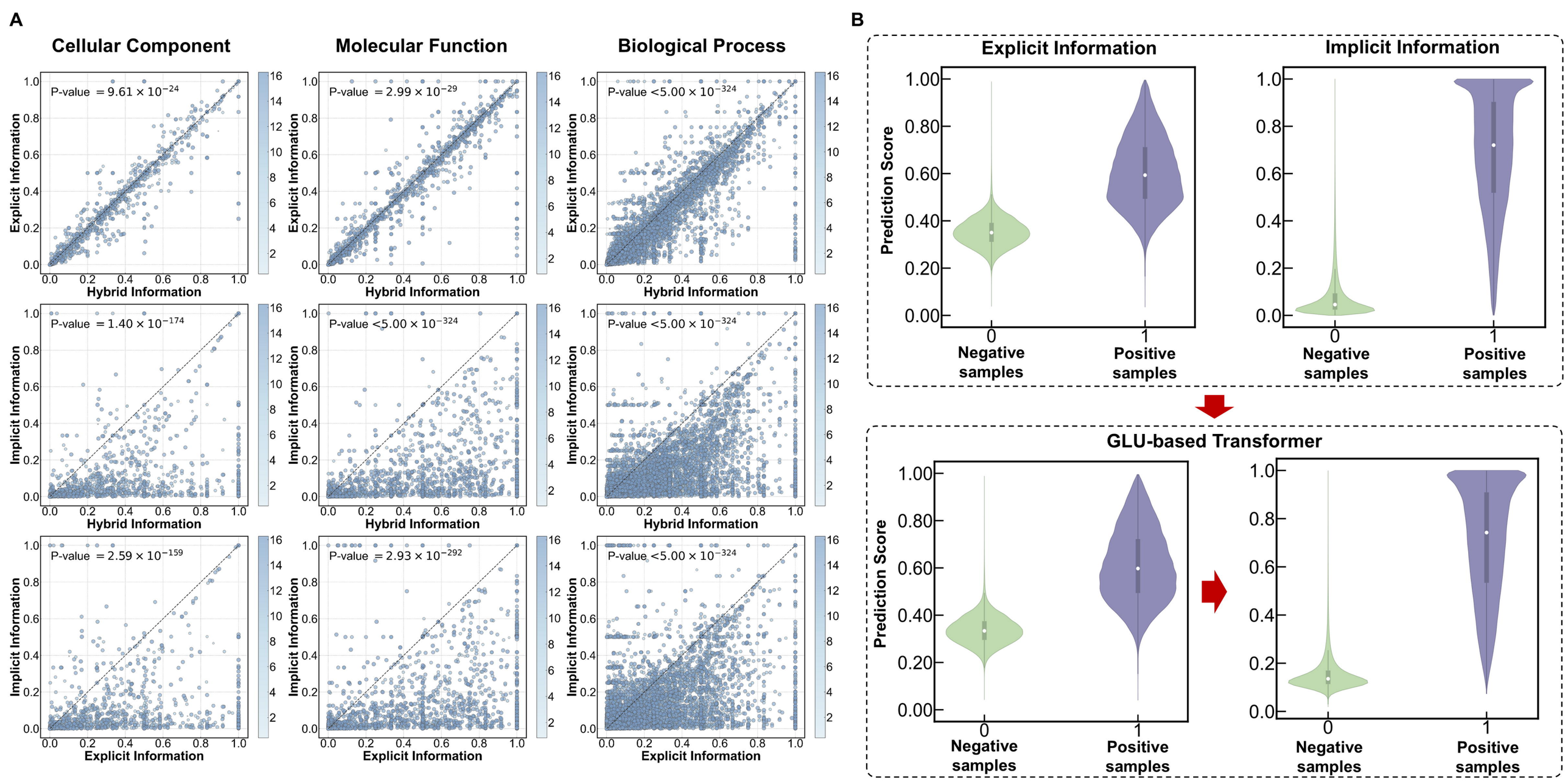
Figure 6. The interpretable representation space learned by ProGO-PSL is capable of capturing protein function information. We analyzed the specificity level imbalance problem via the derived representation of ProGO-PSL. (A) We extracted the trained weights of the last layer of ProGO-PSL and DeepGOPlus and visualized their distributions using the t-SNE algorithm. In the scatter diagram, each point denotes a Gene Ontology (GO) term, where the points colored purple denote highly specific GO terms. (B) We divided the test samples into different groups based on their sequence similarity to the training samples. We employed the Spearman's correlation coefficient to evaluate how closely the predicted results match the actual distribution. Spearman-1 is for the highly specific GO terms and Spearman-2 is for all GO terms. (C) We calculated protein similarity in two ways: one based on the representation of ProGO-PSL, the other based on GO annotations. To compare the two types of protein similarity distributions, we plotted a scatter diagram where each point denotes a protein, and its coordinates represent the similarity score pair. Then, we calculated the rho value to measure how closely the predicted similarity scores match the true ones.

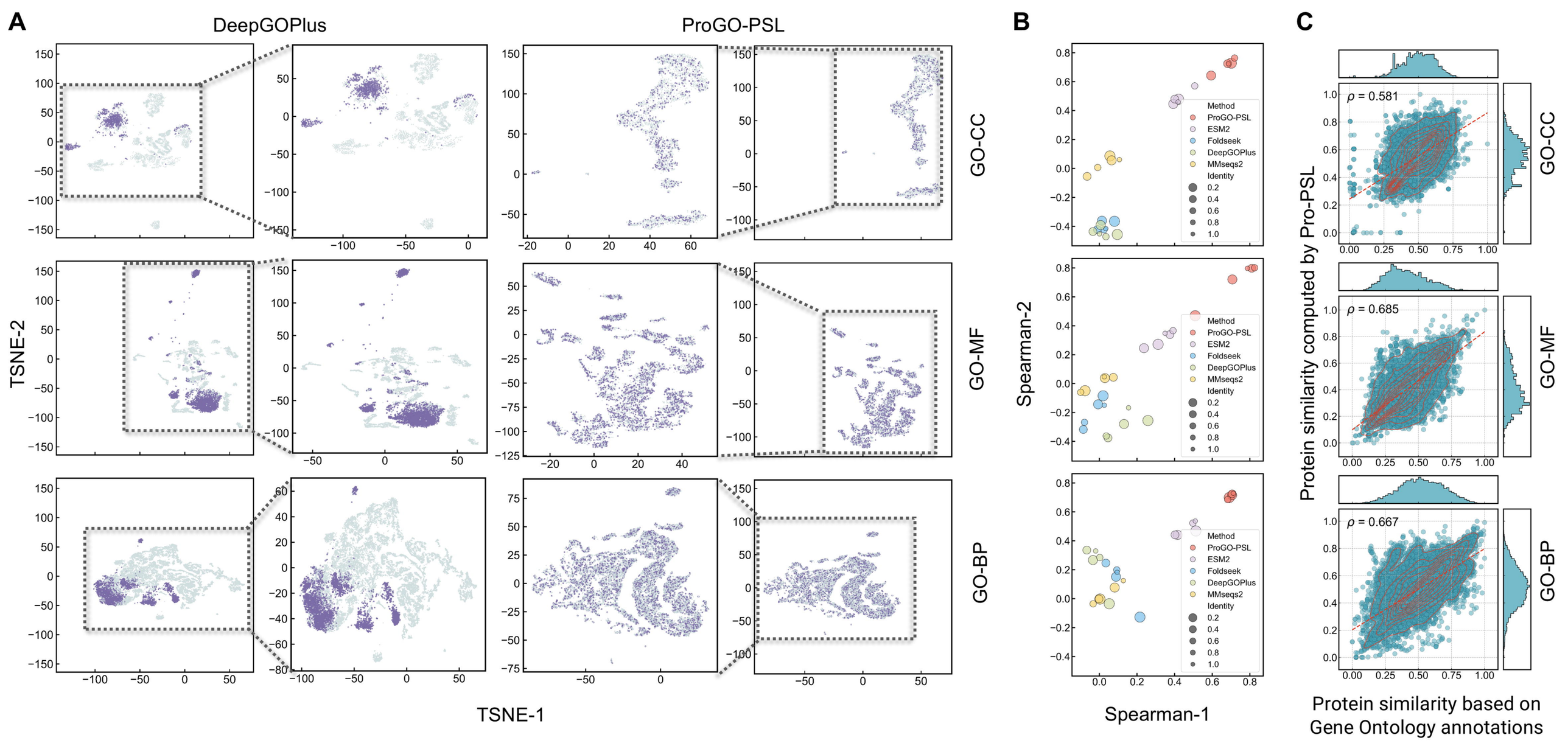

A
Cellular component
Molecular function
Biological process
Protein sequence
Case I
Case II
Case III
Case IV
B
Cellular component
Molecular function
Biological process
Protein sequence
Case I
Case II
Case III
Case IV
C
Case I: High sequence similarity
GO:0005575
GO:0110165
GO:0043226
GO:0043229
GO:0043227
GO:0043231
GO:0005773
GO:0000323
GO:0000322
function similarity
1.000
Case II: High sequence similarity
GO:0008150
GO:0008152
GO:0009987
GO:0050896
GO:0044281
GO:0044237
GO:0006950
GO:0006082
GO:0009611
GO:0044255
GO:0019372
GO:0010597
function similarity
0.279
Case III: Low sequence similarity
GO:0003674
GO:0003824
GO:0016740
GO:0016787
GO:0016757
GO:0016798
GO:0016758
GO:0004553
GO:0046527
GO:0015927
GO:0033831
GO:0004555
function similarity
0.200
Case IV: Low sequence similarity
GO:0008150
GO:0065007
GO:0050789
GO:0009987
GO:0050794
GO:0007165
GO:0035556
GO:0007264
function similarity
1.000
D
Cellular component
Molecular function
Biological process
Protein sequence
Depth / maximum Depth
IC / maximum IC
Lin Similarity (BMA)
Sequence Identity
0.00
0.25
0.50
0.75
1.00


A
Coverage precision
Gene Ontology with IC < 10
Gene Ontology with IC ≥ 10
Information content of Gene Ontology annotation
Coverage precision
Information content of protein sequence
The depth of Gene Ontology
Information content of Gene Ontology annotation
Protein function similarity
Protein sequence similarity
B
Raw sequence
Explicit information
Implicit information
Hybrid information
Dense
Sparse
Related
Specialized
Dense
Related
Generalized
Hybrid
Related
Specialized
Generalized
Predicted score
Negative samples
Positive samples
C
Stage I: RCAN
Embedding (N=21, n_dims=32)
Multiple Sequence Alignment Reconstruction
SpaceToDepth Block (stride=2)
2D Convolution (3x3, stride=1)
2D Convolution (3x3, stride=1)
Fixed Blur Filter (3x3 kernel, stride=2)
GAP
Space To Depth
2D Conv, 1 x 1
BN (leaky relu)
2D Conv, 3 x 3
BN (leaky relu)
Attention layer
2D Conv, 1 x 1
2D BatchNorm
ReLU
Implicit representation
Stage II: CER
InterPro entries
InterPro Resource
IPR009062
IPR009094
IPR000667
IPR000810
Entry encoding
Feature extraction
Stacking
Explicit representation
Stage III: FoR
Dense
Softmax
Self-Attention
Gated Linear Unit
CCO
MFO
BPO
Input
Self-Attention with Gated Linear Unit
GO Prediction


A
IC ≥ 10
IC < 10
All
Fmax
AuPRC
Cellular Component
Molecular Function
Biological Process
ProGO-PSL
DeepGO-SE
DeepGOZero
DPFunc
ESM2
Foldseek
MMseqs2
Diamond
0.72
0.68
0.54
0.74
0.66
0.55
0.74
0.72
0.56
0.80
0.71
0.54
0.81
0.69
0.55
0.75
0.71
0.49
Method
B
IC ≥ 10
IC < 10
All
Fmax
AuPRC
0.72
0.69
0.52
0.74
0.67
0.53
0.74
0.74
0.54
0.71
0.69
0.48
0.73
0.68
0.47
0.75
0.75
0.50
Method


A
Cellular component
Molecular function
Biological process
IC ≥ 10
IC< 10
All
Precision score
Recall score
ProGO-PSL
DeepGO-SE
DeepGO-Zero
DPFunc
ESM-2
B
Fmax score
AuPRC score
(shallow, high)
(shallow, low)
(deep, high)
(deep, low)
(medium, high)
(medium, low)
ProGO-PSL
DeepGO-SE
DeepGOZero
DPFunc
ESM2
C
MCC score
Threshold
Size of annotations
Predicted score
ProGO-PSL threshold
Score Distribution
IC ranges


A
Cellular Component
Molecular Function
Biological Process
P-value = 9.61 × 10−24
P-value = 2.99 × 10−29
P-value <5.00 × 10−324
P-value = 1.40 × 10−174
P-value <5.00 × 10−324
P-value <5.00 × 10−324
P-value = 2.59 × 10−159
P-value = 2.93 × 10−292
P-value <5.00 × 10−324
Explicit Information
Implicit Information
Hybrid Information
Explicit Information
B
Explicit Information
Implicit Information
Prediction Score
Negative samples
Positive samples
GLU-based Transformer


A
DeepGOPlus
ProGO-PSL
GO-CC
GO-MF
GO-BP
TSNE-2
TSNE-1
B
Spearman-2
Spearman-1
Method
ProGO-PSL
ESM2
Foldseek
DeepGOPlus
MMseqs2
Identity
0.2
0.4
0.6
0.8
1.0
C
Protein similarity computed by Pro-PSL
Protein similarity based on Gene Ontology annotations
ρ = 0.581
ρ = 0.685
ρ = 0.667