Language-Encoded Structural Topology Enables Generalizable Foundation Models for Graph-Structured Data

Sakib Mostafa[1], James Zou[2,3], Ash A. Alizadeh[4], Maximilian Diehn[1,5,6], Lei Xing[1,7,8,*], and Md Tauhidul Islam[1,*]

[1]Department of Radiation Oncology, Stanford University, Stanford, CA, USA
[2]Department of Computer Sciences, Stanford University, Stanford, CA, USA
[3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[4]Department of Medicine, Division of Oncology, Stanford University, Stanford, CA, USA
[5]Stanford Cancer Institute, Stanford University, Stanford, California, USA
[6]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA
[7]Institute of Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA
[8]Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[*]Correspondence: tauhid@stanford.edu & lei@stanford.edu

ABSTRACT

Graphs are a central representation in biomedical research, capturing molecular interaction networks, gene regulatory circuits, cell–cell communication maps, and knowledge graphs. Despite their importance, currently there is no broadly reusable foundation model available for graph analysis comparable to those that have transformed language and vision. Existing graph neural networks are typically trained on a single dataset and learn representations specific only to that graph's node features, topology, and label space, limiting their ability to transfer across domains. This lack of generalization is particularly critical in biology and medicine, where networks vary substantially across cohorts, assays, and institutions. Here we introduce a graph foundation model designed to learn transferable topological representations that are not specific to specific node identities or feature schemes. Our approach derives feature-agnostic graph properties, including degree statistics, centrality measures, community structure indicators, and diffusion-based signatures, and encodes them as topological prompts. The resulting prompt embeddings are combined with node features and processed by a message-passing deep learning backbone to learn graph neighborhood embeddings. During pretraining, contrastive alignment maps the prompt embeddings and graph neighborhood embeddings into a common representation space across graphs, bringing nodes with similar topological roles closer and separating nodes with different structural roles. The proposed model is pretrained once on diverse heterogeneous non-biological graphs comprising approximately 279,000 nodes and 4.4 million edge-index entries, and is evaluated on 18 held-out biological graphs and 6 non-biological graphs, spanning 37 downstream evaluation tasks with minimal adaptation. Across multiple benchmarks, our pretrained model matches or exceeds strong supervised baselines while demonstrating superior zero-shot and few-shot generalization on held-out graphs. On the SagePPI benchmark, supervised fine-tuning of our pretrained model achieves a mean ROC-AUC of 95.5%, a gain of 21.8% over the best supervised graph neural networks. Consistent advantages are observed across diverse biological network benchmarks spanning protein interaction networks, GO annotation prediction, and protein fold classification. These results establish topological graph properties as a transferable inductive structure for reusable graph models that generalize across biological and non-biological domains.

Introduction

Graphs are widely used to represent relational systems in biomedical science, including molecular and protein–protein interaction networks[1,2], gene regulatory circuits[3], and cell–cell communication maps[4]. They are also the inherent representation for biomedical knowledge graphs that link genes, diseases, drugs, and phenotypes[5]. This wide applicability of graph representations makes graph learning critically important for both biological discovery and clinical translation. In recent years, deep learning has transformed representation learning in domains with fixed input formats, most notably language and vision, where large pretrained models can be reused across tasks with minimal adaptation[6–8]. However, despite major advances in graph neural networks, there is still no broadly reusable foundation model that generalizes reliably across graphs from different domains. This challenge is particularly pronounced in biology and medicine, where graph structure can vary across cohorts, assays, and institutions, causing models trained on one graph to transfer poorly to another.

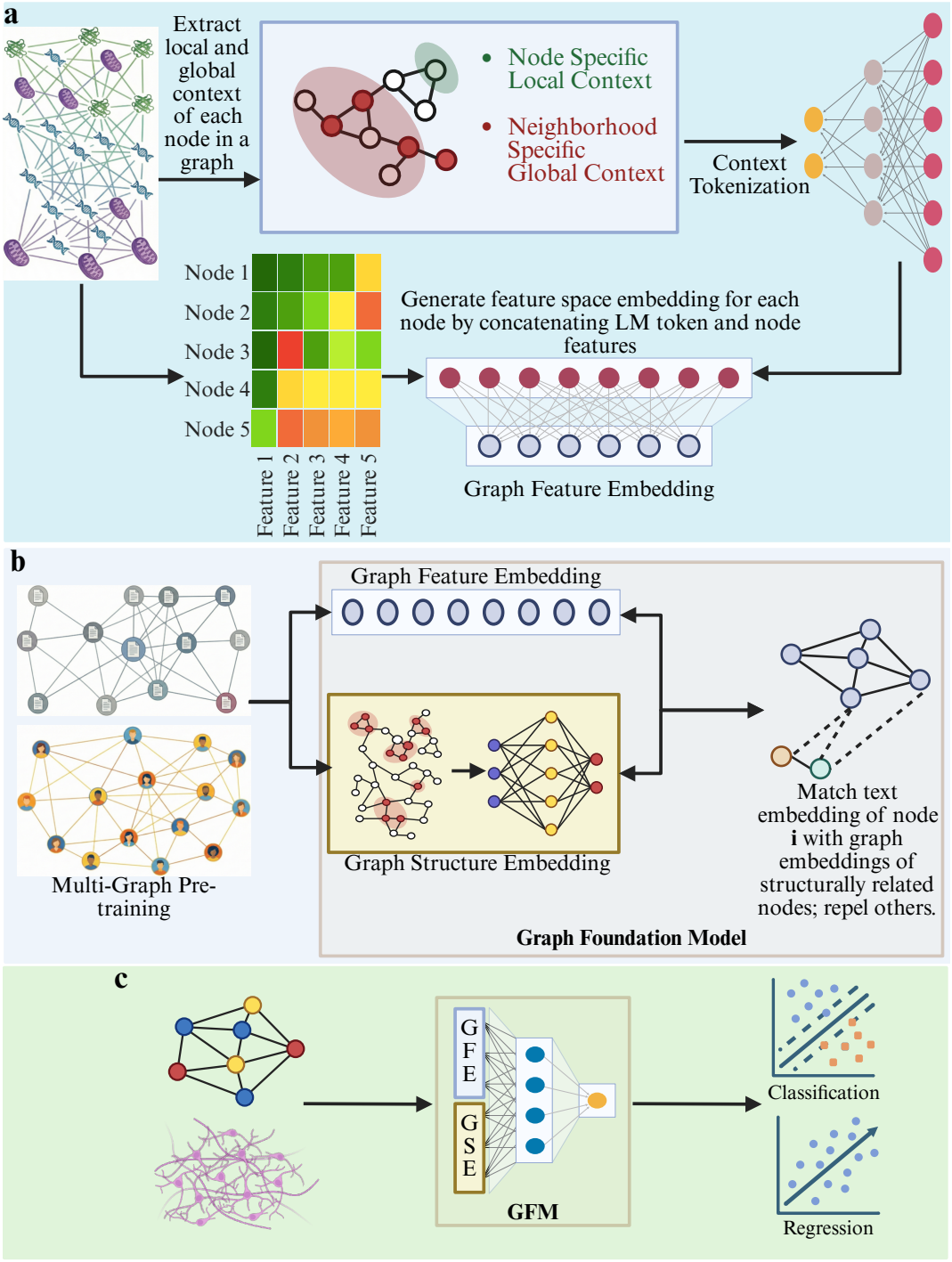

Figure 1. Architecture of the proposed Graph Foundation Model. a, Topological prompt construction. For each node, local context descriptors—degree, clustering coefficient, k-core number, ego-network statistics, and PageRank—are combined with global context indicators derived from two community detection algorithms, Label Propagation and SCoDA, each contributing community identity, community size, and intra-community density. These descriptors are assembled into a natural-language node profile and tokenized by a miniature language model (MiniLM) to produce a context embedding. Node features are projected into a matching space and concatenated with this embedding to form the Graph Feature Embedding (GFE), a representation that encodes both attribute content and topological role without relying on dataset-specific feature semantics. **b, Multi-graph pretraining**. A deep learning backbone ingests multiple heterogeneous graphs simultaneously, producing a Graph Structure Embedding (GSE) for every node via neighborhood aggregation. Contrastive alignment is applied between the GFE and the GSE: the text embedding of node i is matched against the graph embeddings of structurally related nodes and repelled from structurally dissimilar nodes, training the backbone to map diverse graphs into a shared topological representation space. **c, Downstream deployment**. The frozen or fine-tuned GFM generates GFE and GSE representations for nodes in an unseen graph. These are concatenated and passed to a lightweight task-specific head for node-level classification or regression, requiring minimal labeled data on the target graph.

Most existing graph neural networks are trained on a single dataset and learn representations that are specific to that graph's node features, topology, and label space[9]. Common message-passing architectures, including Graph Convolutional Networks (GCN)[10], Graph Isomorphism Networks (GIN)[11], Graph Attention Networks (GAT)[12], and GraphSAGE[13], can be used on semi-supervised learning and node classification tasks when training and testing occur within the same graph. However, the learned parameters typically reflect the statistics and semantics of the dataset the networks were trained on. Recent self-supervised approaches, including Deep Graph Infomax (DGI)[14], GRACE[15], and BGRL[16], have the same limitations. They define objectives on a single graph and learn representations that are most reliable within that same structural and feature distribution. As a result, graph representation learning remains dataset-specific in practice. Models are often retrained from scratch for each new graph, and knowledge learned from one dataset is not consistently carried forward to another[17], even when the graphs describe related biological systems.

The generalizability limitation of graph neural networks stems from properties that are inherent to graphs. Unlike images or text, graphs do not have a canonical input space[18–20]. Node features can take many forms, such as bag-of-words vectors, molecular fingerprints, continuous attributes, or learned embeddings, and their meaning varies widely across applications. The meaning of edges also changes across domains. An edge may represent a physical interaction, a regulatory relationship, a citation, or a transactional link. This variability makes it difficult to define a shared representation that works across graphs from various domains. There are also architectural constraints. Most graph models rely on local message passing, where each layer aggregates information from a node's neighbors. This works well for local structure, but long-range dependencies require deeper networks. As depth increases, node representations can become overly similar[21], and information from distant regions of the graph can be compressed into small vectors. These effects weaken the model's ability to preserve global structural signals consistently. Pretraining alone does not resolve this problem when the objective is still defined within a single graph. In that setting, the model learns invariances that reflect one graph's feature distribution, structural statistics, and label semantics, rather than graph properties that remain comparable across datasets. As a result, existing pretraining objectives do not transfer reliably under large shifts in node feature space, graph structure, or label semantics. Together, these properties for graph data and analysis methods make cross-graph generalization a fundamental challenge rather than an incidental limitation.

For a reusable foundation model, the ideal representation must rely on components of graph organization that are intrinsic and broadly available, rather than being tied to dataset-specific features or labels. Graph topological properties provide such a basis which can be defined without reference to node semantics. These include degree statistics, centrality measures, community structure, diffusion statistics, and higher-order connectivity patterns[22]. They can be computed for any graph and can be compared across graphs, even when the underlying node attributes differ. If these topological signals can be expressed through a common interface, then heterogeneous graphs may be mapped into a shared embedding space without requiring common node identities or a shared feature schema. This provides a practical route to cross-domain learning in settings where raw features are incompatible, but topology remains meaningful.

Here, to address the limitations of existing graph analysis methods and utilize the generalizable graph topological properties, we introduce a graph foundation model that learns transferable topological representations not tied to any single graph's node identities or feature schema. Our approach converts feature-agnostic graph properties into structured prompts at the node level. These prompts summarize a node's position and role in the graph using degree statistics, centrality measures, community indicators, and diffusion-based signatures, along with related topological descriptors such as clustering and core structure. When textual node descriptors are available, they are included as additional context. We integrate these prompts with a lightweight message-passing backbone. The prompts provide an explicit topological description, while message passing captures local relational patterns that depend on connectivity. Contrastive training aligns the topology-derived prompt embeddings with the message-passing neighborhood embeddings. It brings nodes with similar topological roles closer in the shared embedding space and separates nodes with dissimilar structural roles. This objective does not require shared node identities or common feature schemas, making it applicable to any graph for which topological descriptors can be computed. The backbone thereby learns representations that remain meaningful across graphs even with fundamentally different node feature spaces. Because the topological interface is defined from topology, the model can align graphs without requiring common node identities, shared attributes, or shared label sets, which are rarely available across biomedical datasets.

We pretrain the proposed model once on a heterogeneous collection of non-biological graphs and then reuse the frozen backbone on unseen datasets for different biomedical applications. We evaluate our approach on four held-out biological network benchmarks: SagePPI, a human protein–protein interaction network with GO biological process annotations[13]; OGBN-Proteins, a large-scale STRING-derived interaction network from the Open Graph Benchmark with species-stratified evaluation[23,24]; StringGO, a STRING network with GO annotations spanning three ontologies evaluated using amino acid composition features[24,25]; and Fold-PPI, a collection of 144 tissue-specific interaction networks with protein structural fold class labels drawn from SCOP[26,27]. Across these benchmarks, our pretrained model matches or exceeds strong supervised baselines in zero-shot mode and shows substantially improved generalization when labels are limited. These findings indicate that pretraining on heterogeneous graphs using our approach produces representations that capture biologically meaningful

organization without requiring task-specific training on the target network.

Results

Topology-based pretraining produces transferable functional representations in a human protein interaction network

We first evaluated the GFM on the SagePPI protein–protein interaction benchmark[13], which comprises experimentally supported physical interactions among human proteins annotated with 121 Gene Ontology (GO)[28] biological process labels drawn from the MSigDB GO:BP 2021 collection[29]. Because individual proteins can participate in multiple biological processes, the benchmark is formulated as multilabel node classification, where each protein is independently scored against all 121 process labels. This evaluation assesses whether topology-derived embeddings learned from non-biological pretraining graphs can organize proteins according to shared biological function in a held-out interaction network. The SagePPI benchmark therefore provides a stringent cross-domain generalization setting: the network was not included during pretraining, and downstream performance is evaluated on held-out proteins from the benchmark split.

The pretrained GFM in zero-shot mode, using a frozen backbone with a lightweight linear probe, achieved a mean ROC-AUC of 76.1% and an accuracy of 72.2% across all 121 GO process labels (Fig. 2e). This exceeded the strongest supervised baseline, GIN[11], which reached a mean ROC-AUC of 73.7% under the same evaluation protocol despite being trained directly on SagePPI node features and topology. GraphSAGE[13] and GAT[12] reached 69.7% and 67.1% respectively, while GCN[10] achieved 57.1%. When the GFM backbone was fine-tuned on SagePPI with full supervision, mean ROC-AUC increased substantially to 95.5% and accuracy to 84.1%, representing a gain of 21.8 percentage points over the best supervised baseline. Macro-averaged ROC curves across all 121 labels confirm that both GFM variants dominate the baselines at every operating threshold, with the separation most pronounced at low false positive rates where precision requirements are most stringent (Fig. 2d; Supplementary Fig. S1).

To assess whether this advantage reflects a general capacity for rapid adaptation rather than performance that requires full supervision, we evaluated each model under a few-shot protocol in which a logistic probe was trained using only K labeled examples per GO process, with $K \in \{1, 5, 10, 20\}$ (Fig. 2a). At $K = 1$, the GFM zero-shot embeddings achieved a mean ROC-AUC of 0.567, while all supervised baselines clustered between 0.505 and 0.515 at this extreme label scarcity. At $K = 20$, GFM zero-shot reached 0.690 and GFM trained reached 0.661, while the best-performing baseline, GraphSAGE, reached only 0.583. The pretrained GFM thus achieves at $K = 1$ a level of performance that supervised baselines approach only after accumulating 20 labeled examples per GO process. GFM trained embeddings showed a complementary trajectory: starting near baseline level at $K = 1$ and rising steeply as labels accumulate, consistent with the interpretation that fine-tuned representations encode more discriminative task-specific structure that benefits substantially from downstream supervision. Across the entire few-shot range, both GFM variants outperformed all supervised baselines, indicating that pretraining instills transferable functional structure that task-specific training from scratch does not recover even with equivalent labeled data.

Beyond discriminative performance, we examined whether GFM embeddings organize proteins according to biologically coherent functional neighborhoods. For six representative GO processes spanning distinct biological programs—DNA repair, cell cycle regulation, Pol-II transcription, cell proliferation regulation, negative regulation of transcription, and nervous system development—we computed the same-label fraction among the k nearest neighbors in each model's embedding space for $k \in \{5, 10, 20, 30, 50, 75, 100\}$ (Supplementary Figs. S2, S3). GFM zero-shot and GFM trained embeddings consistently maintained higher same-label fractions at all values of k relative to all baselines, with the gap most pronounced at small k where local neighborhood purity is most informative about embedding organization. This advantage was reproducible across all six processes without exception, indicating that functional cohesion in the GFM embedding space is not confined to a single annotation class but reflects a broadly organized representation of protein function.

To characterize the joint functional structure of the embedding space, we computed a co-enrichment matrix across the 20 most populated GO processes, quantifying the mean fraction of annotated neighbors that a protein from one process accumulates within a 25-nearest-neighbor embedding neighborhood defined by GFM trained representations (Fig. 2c). Hierarchical clustering of this matrix revealed a block structure in which GO processes with established biological relationships clustered together without prior knowledge of their groupings. Transcription-related processes—Pol-II transcription, positive and negative regulation of transcription, and regulation of gene expression—formed a coherent cluster with elevated inter-process enrichment, consistent with their shared mechanistic basis in gene regulatory circuits. Immune and cytokine-related processes—neutrophil immunity, neutrophil activation, cytokine signaling, and cellular cytokine response—formed a separate block, reflecting their shared involvement in inflammatory programs. These patterns indicate that the GFM embedding space preserves not only within-class functional cohesion but also the higher-order relational structure among biological processes, a property that emerges from topology-based pretraining rather than from any explicit annotation of process relationships.

We further examined whether the embedding density of a protein in the GFM zero-shot space reflects its degree of multifunctionality, defined as the number of GO process annotations assigned to that protein (Fig. 2b). Spearman rank

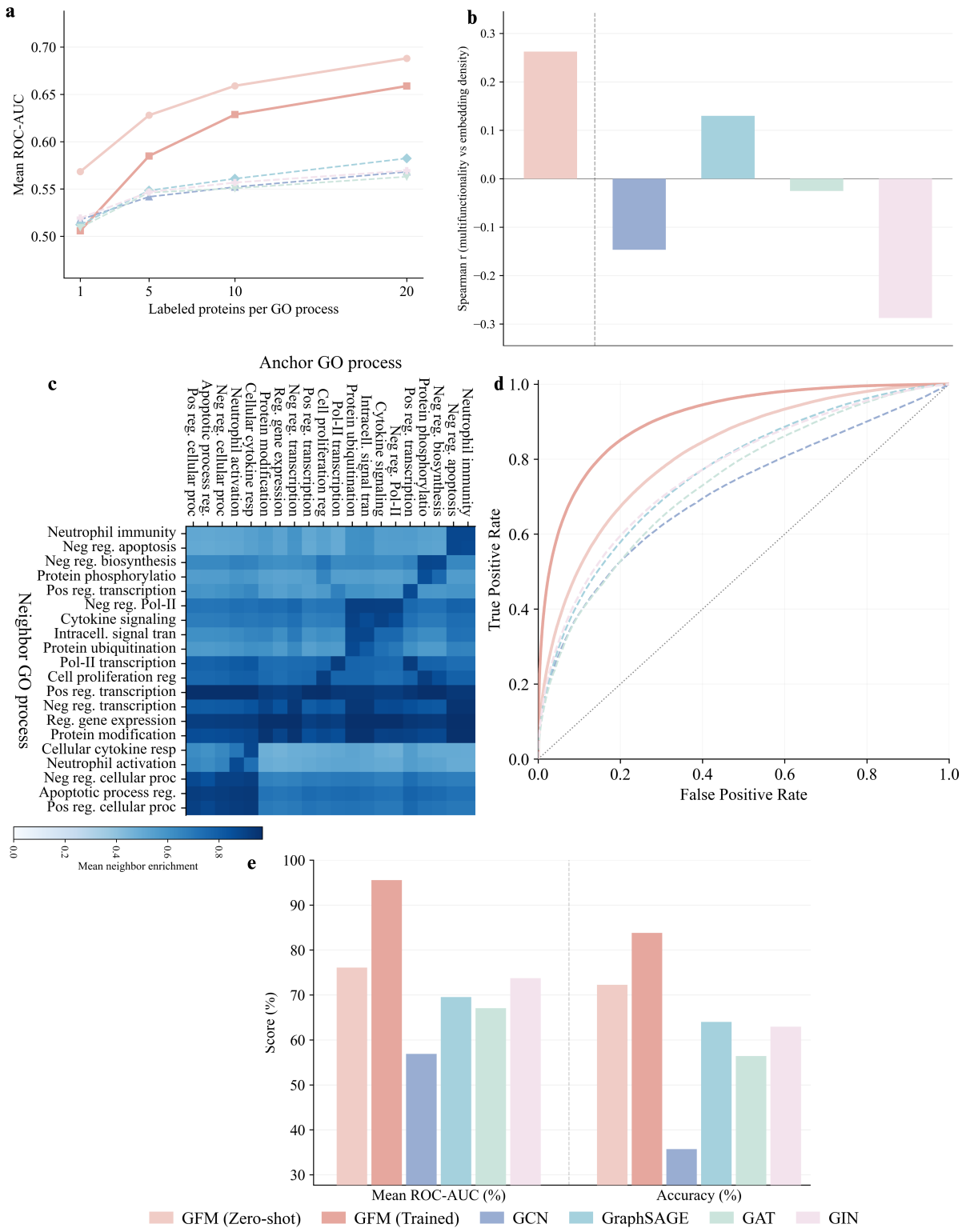

Figure 2. The pretrained GFM encodes protein functional organization in SagePPI and generalizes with minimal supervision. a, Few-shot generalization curves. Mean ROC-AUC is shown as a function of the number of labeled proteins per GO process, evaluated on held-out test proteins using a logistic probe trained on K positive and K negative examples per label ($K \in \{1, 5, 10, 20\}$). Solid lines denote GFM variants; dashed lines denote supervised baselines. **b, Multifunctionality–embedding density correspondence**. Bars show the Spearman rank correlation between the number of GO process annotations per protein and the local embedding density computed from the 15 nearest neighbors in each model's embedding space. The dashed vertical line separates the pretrained GFM zero-shot model (left) from supervised GNN baselines (right). Positive values indicate that highly annotated proteins occupy denser embedding neighborhoods. **c, Cross-functional co-enrichment in GFM trained embeddings**. Each cell reports the mean fraction of GO-annotated neighbors that a protein from one anchor process accumulates within a 25-nearest-neighbor embedding neighborhood annotated by another process. Rows and columns are ordered by hierarchical clustering of the symmetrized co-enrichment matrix. Darker entries indicate higher cross-process neighborhood sharing. **d, Macro-averaged ROC curves across all 121 GO process** labels. Each curve

correlation between per-protein label count and local embedding density yielded $r = +0.263$ for GFM zero-shot. In contrast, GCN produced $r = -0.131$, GraphSAGE $r = +0.130$, GAT $r = -0.025$, and GIN $r = -0.290$. The strongly negative correlation in GIN embeddings indicates that supervised training under this architecture places highly multifunctional proteins at the periphery of the embedding space, which is inconsistent with what protein interaction network topology would predict. The positive correlation in GFM zero-shot embeddings is consistent with the expectation that proteins participating in many biological processes occupy structurally central positions in the interaction network[30] and should therefore be embedded in densely populated regions of a topology-informed representation space. This property emerges without any supervision from functional labels and therefore reflects structure that pretraining captures directly from graph topology.

Taken together, these results establish that the pretrained GFM encodes transferable topological representations that organize the SagePPI interaction network according to protein functional programs. In zero-shot mode, the frozen backbone with a linear probe surpasses the best supervised GNN baseline, and in the few-shot regime it achieves with a single labeled example per process what supervised models require 20 labeled examples to approach. The embedding space preserves functional cohesion across diverse GO biological processes, recovers known relationships among annotation classes through its co-enrichment structure, and places multifunctional proteins in topologically central positions. These findings indicate that topology-based pretraining encodes functional organization at a level that does not depend on the specific proteins, species, or annotation vocabulary present during training.

Cross-species generalization scales with GO annotation specificity in OGBN-Proteins

We next evaluated the GFM on the OGBN-Proteins benchmark from the Open Graph Benchmark[23], a large-scale protein–protein interaction network derived from the STRING database[24] with 112 GO biological process annotations per node as multilabel targets. The benchmark uses a species-stratified split in which test proteins come from species not represented during training. This makes cross-species generalization a requirement for strong test performance. Models that rely primarily on species-specific graph patterns are therefore expected to generalize poorly. This benchmark evaluates whether the topology-based pretraining strategy in our approach produces embeddings that transfer across biological contexts.

The pretrained GFM in zero-shot mode achieved a mean ROC-AUC of 71.0% across all 112 GO labels (Fig. 3b). This exceeded GCN (56.5%), GraphSAGE (50.9%), GAT (47.0%), and GIN (64.6%), the strongest supervised baseline. What makes this margin meaningful is that GIN operates on confidence-weighted STRING edges during training[24], providing explicit interaction reliability information that the GFM never receives. When the GFM backbone was fine-tuned with full supervision, mean ROC-AUC increased to 77.5% with an accuracy of 81.1%, a gain of 12.9 percentage points over GIN. Macro-averaged ROC curves confirm that both GFM variants dominate the baselines at every operating threshold (Fig. 3a; Supplementary Fig. S4). The few-shot evaluation reinforces these findings (Fig. 3d). At $K = 1$ labeled protein per GO process, GFM zero-shot achieved a mean ROC-AUC of 0.620. All supervised baselines fell between 0.495 and 0.540 at this shot count despite having been trained on the full graph with complete label supervision. At $K = 20$, GFM zero-shot reached 0.738 while the best baseline reached 0.646. The GFM few-shot curve rises steeply from $K = 1$ and continues improving, indicating that pretrained topological embeddings provide a strong initialization that benefits from small amounts of additional supervision. Baseline trajectories remain comparatively flat across the same range.

The most distinctive result in this dataset concerns GO hierarchy depth (Fig. 3e). Shallow GO terms are broad and annotate many proteins[25]. Deep terms are specific and annotate far fewer. For GFM trained, mean ROC-AUC increased monotonically from 75.4% at shallow terms to 76.7% at medium depth and 80.5% at deep terms. GIN showed no consistent directional trend, producing 61.7% at shallow, 61.4% at medium, and 67.5% at deep. GFM zero-shot similarly improved with depth, from 70.0% at shallow to 71.4% at deep. The GFM widens its advantage over baselines precisely where the task is hardest. Deep GO terms require integrating information across long network paths and community boundaries to identify specific functional roles. Topology-based pretraining captures these signals through diffusion statistics, k-core decomposition[31], and multi-scale community indicators[32,33]. Supervised models trained on a single graph do not acquire this capacity from labeled data alone.

The multifunctionality analysis replicates the SagePPI result on a graph with a different annotation system and edge weighting scheme (Fig. 3c). GFM zero-shot produced a Spearman rank correlation of $r = +0.105$ between per-protein GO annotation count and local embedding density. GCN produced $r = +0.052$, GAT $r = +0.079$, GraphSAGE $r = -0.155$, and GIN $r = -0.020$. The positive correlation appears in GFM embeddings on both SagePPI and OGBN-Proteins, across different databases and annotation systems. This consistency supports the interpretation that pretraining from topology places topologically central, highly annotated proteins in dense embedding regions as a general property of the learned representation space.

Co-enrichment analysis of the GFM trained embedding space shows that biological processes are organized into functionally coherent blocks without any explicit knowledge of process relationships (Supplementary Fig. S6). An immune module comprising regulation of immune system process, antigen processing and presentation, acute inflammatory response, immune system process, and immune effector process shows enrichment ratios substantially above 1.0 within the block. A cytokine

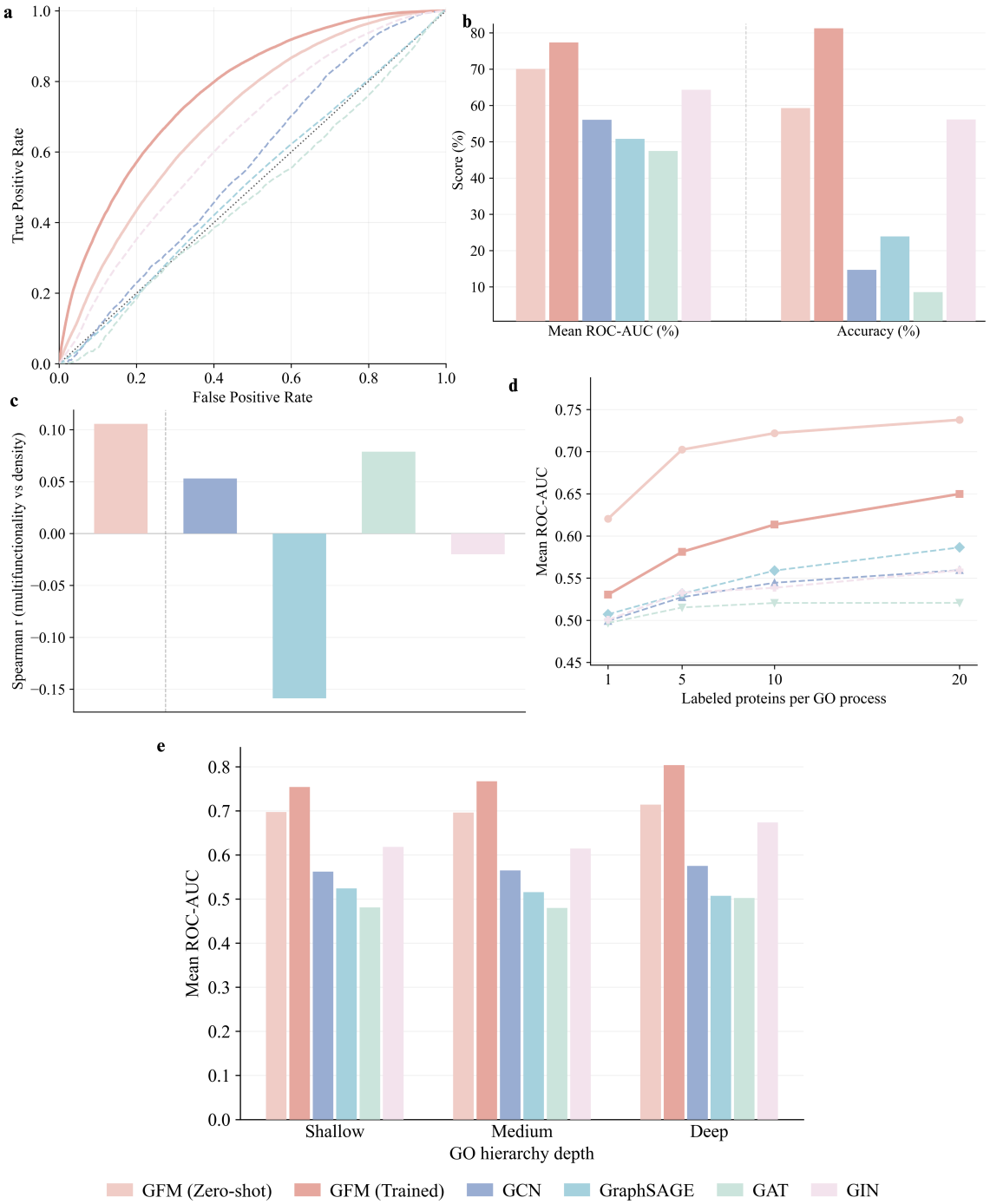

Figure 3. The pretrained GFM generalizes across the OGBN-Proteins benchmark and captures GO hierarchy structure. a, Macro-averaged ROC curves across all GO term labels for GFM zero-shot, GFM trained, and four supervised GNN baselines (GCN, GraphSAGE, GAT, GIN). Solid lines denote GFM variants; dashed lines denote supervised baselines. The dotted diagonal denotes random performance. **b, Summary classification performance**. Mean ROC-AUC and accuracy are shown for all six models. The dashed vertical line separates the two reported metrics. All models use the standard OGBN-Proteins train/validation/test split. **c, Multifunctionality–embedding density correspondence**. Bars show the Spearman rank correlation between the number of GO annotations per protein and local embedding density computed from the 15 nearest neighbors. The dashed vertical line separates the pretrained GFM zero-shot model from supervised baselines. Positive values indicate that highly annotated proteins occupy denser embedding neighborhoods. **d, Few-shot generalization curves**. Mean ROC-AUC is shown as a function of the number of labeled proteins per GO process ($K \in \{1, 5, 10, 20\}$), evaluated on held-out test proteins using a logistic probe trained on K positive and K negative examples per label. Solid lines denote GFM variants; dashed lines denote supervised baselines. **e, GO hierarchy depth stratification**. Mean ROC-AUC is shown separately for GO terms at shallow, medium, and deep levels of the GO biological process hierarchy. GFM trained performance increases with hierarchy depth while baseline performance remains flat or declines.

signaling module covering positive regulation of cytokine production, regulation of cytokine production, cytokine production, and cell activation shows strong internal enrichment and cross-enrichment with the immune block. This is consistent with the known coupling between cytokine-mediated signaling and acute inflammatory programs[34]. Same-label enrichment curves across six GO annotation categories confirm that GFM embeddings maintain higher functional neighborhood purity at all values of k, with the advantage widening for the more specific metabolic process terms relative to broad categories such as molecular function (Supplementary Fig. S5).

The OGBN-Proteins results strengthen the core claim of this work in a specific way. The species-stratified split means that the test proteins come from biological contexts the model has never seen. The GFM zero-shot model still exceeds three supervised baselines and narrows the gap with GIN to 6.4 percentage points under these conditions. Its advantage over baselines grows with GO hierarchy depth rather than shrinking, which is the opposite of what a model relying on local pattern memorization would produce. These findings indicate that structural pretraining encodes functional organization at a level that does not depend on the specific proteins, species, or annotation vocabulary present during training.

Topological pretraining recovers subcellular compartment organization from amino acid sequence networks in StringGO

In SagePPI and OGBN-Proteins dataset we evaluated transfer on protein interaction benchmarks with node representations derived from gene-set, interaction, or edge-associated information. We next evaluated our methods on the StringGO dataset to determine whether the pretrained representations remain effective when node attributes instead encode primary sequence composition. StringGO is a human protein interaction network derived from the STRING database[24] with GO annotations evaluated separately across three ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)[25]. Molecular Function annotations describe the biochemical activities of individual proteins. Biological Process annotations describe the larger biological programs in which proteins participate. Cellular Component annotations describe the subcellular locations where proteins are active. Each ontology defines an independent multilabel classification task on the same underlying interaction graph. Node features are 20-dimensional amino acid composition vectors encoding local sequence triplet statistics, which carry primary sequence properties rather than expression or pathway information. This evaluation tests whether topology-based pretraining remains transferable when target-graph node features represent sequence composition rather than interaction-derived or expression-associated information.

Across Molecular Function and Cellular Component, the pretrained GFM in zero-shot mode achieved mean ROC-AUC values of 81.4% and 81.1% respectively, exceeding the strongest supervised baseline, GIN, which reached 74.7% on Molecular Function and 74.9% on Cellular Component (Fig. 4a,b). GCN, GraphSAGE, and GAT fell substantially further behind, reaching between 42.2% and 53.4% across the two ontologies. When the GFM backbone was fine-tuned with full supervision, mean ROC-AUC increased to 86.3% on Molecular Function and 87.4% on Cellular Component, representing gains of 11.6 and 12.5 percentage points respectively over GIN. Macro-averaged ROC curves confirm the consistent separation between GFM variants and all baselines across the full threshold range for both Molecular Function and Cellular Component (Fig. 4c,d). The margin over GCN, GraphSAGE, and GAT is particularly large, exceeding 30% in mean ROC-AUC, which indicates that standard message-passing architectures struggle to extract functional signal from amino acid composition features alone without the topological scaffolding that pretraining provides. On the Biological Process ontology, the GFM zero-shot model achieved a mean ROC-AUC of 80.5%, which again exceeded all supervised baselines including GIN (69.3%), GCN (44.9%), GraphSAGE (50.6%), and GAT (51.4%) (Supplementary Fig. S7). The zero-shot advantage on Biological Process is therefore consistent with Molecular Function and Cellular Component.

GFM embeddings also retain functional structure under amino acid feature perturbation. At 50% noise applied to the 20-dimensional amino acid composition input, GFM zero-shot maintained a mean ROC-AUC of 0.791 on Molecular Function and 0.791 on Cellular Component, representing a decline of only 2.3 and 2.0 percentage points from the clean-feature baseline respectively (Supplementary Fig. S8). This stability reflects the fact that the GFM embedding is primarily driven by the topology tokens, which encode community membership, diffusion statistics, and local connectivity and are not affected by amino acid feature perturbation.

Co-enrichment analysis of the GFM trained embedding space reveals biologically coherent term groupings for both Molecular Function and Cellular Component (Fig. 4e,f). In the Molecular Function heatmap, a cluster of DNA-binding and RNA polymerase II activity terms shows elevated mutual enrichment, consistent with the shared mechanistic role of these factors in transcription initiation and elongation[35]. The olfactory receptor activity term shows strong self-enrichment and high enrichment with the broader protein binding category, reflecting the well-known protein-protein interaction network of olfactory receptors and their shared structural scaffold[36]. The Cellular Component heatmap shows three clearly resolved modules. A mitochondrial module groups mitochondrial matrix, mitochondrion, and mitochondrial inner membrane together with high mutual enrichment. A nuclear and cytoplasmic module groups nucleus, nucleoplasm, cytosol, cytoplasm, and chromatin, capturing the shared spatial occupancy of nuclear proteins within the cell. A membrane module groups plasma membrane

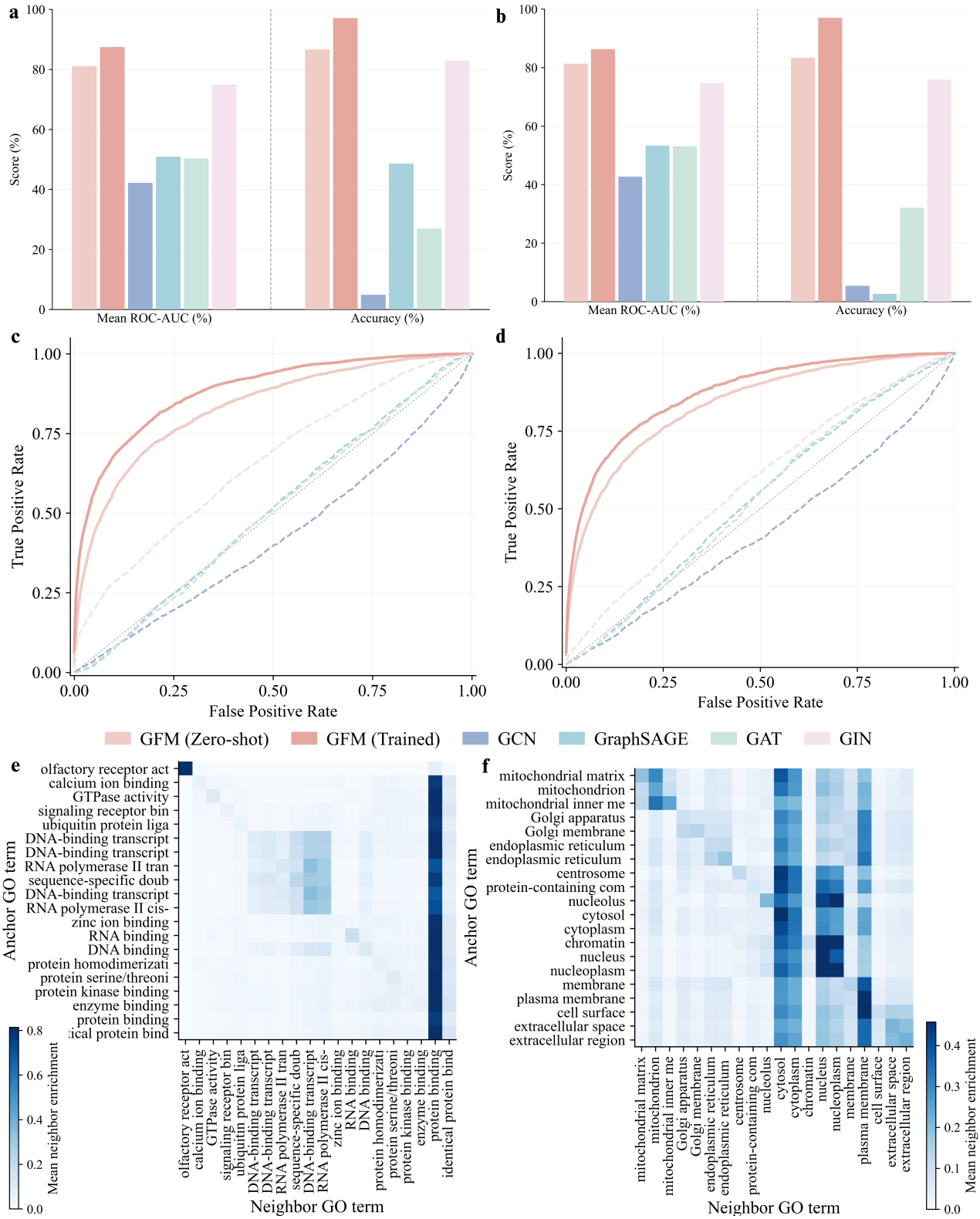

Figure 4. The pretrained GFM generalizes across GO ontologies in the STRING interaction network. a, Mean ROC-AUC and accuracy for Molecular Function (MF) GO term prediction. All six models are compared under the same train/validation/test split on the StringGO network. **b, Mean ROC-AUC and accuracy for Cellular Component (CC) GO term prediction under identical conditions**. **c, Macro-averaged ROC curves for Molecular Function**. Solid lines denote GFM variants; dashed lines denote supervised baselines. The dotted diagonal denotes random performance. **d, Macro-averaged ROC curves for Cellular Component**. Same protocol as c. **e, Cross-functional co-enrichment in GFM trained embeddings for Molecular Function**. Each cell reports the mean fraction of GO-term annotated neighbors that a protein from one anchor Molecular Function term accumulates within a 25-nearest-neighbor embedding neighborhood annotated by another term. Rows and columns are ordered by hierarchical clustering of the symmetrized co-enrichment matrix. **f, Cross-functional co-enrichment for Cellular Component**. Same protocol as e. Three distinct spatial modules are visible: a mitochondrial module, a nuclear and cytoplasmic module, and a membrane module.

and membrane, which is consistent with their overlapping protein populations. These three modules correspond directly to the major spatial compartments of eukaryotic cells and emerge from network topology alone without any explicit annotation of compartment membership provided to the model. The Biological Process co-enrichment matrix shows similar biological coherence, with distinct immune, signaling, and transcriptional clusters visible (Supplementary Fig. S9). This confirms that the zero-shot embedding space organizes Biological Process annotations correctly, although the supervised head does not improve on this organization under the current fine-tuning procedure. Cross-ontology sensitivity analysis further shows that GFM Trained embeddings place proteins sharing the same dominant GO term in closer neighborhoods than any baseline, consistently across all three ontologies and all values of k from 25 to 100 (Supplementary Fig. S10).

The observed results establish a consistent pattern across three datasets with different node feature modalities: expression-derived features in SagePPI, confidence-weighted interaction features in OGBN-Proteins, and amino acid composition in StringGO. In all three cases the pretrained backbone transfers to the target graph without seeing it during training.

Pretrained topological embeddings outperform supervised baselines on entirely unseen protein fold classes in Fold-PPI

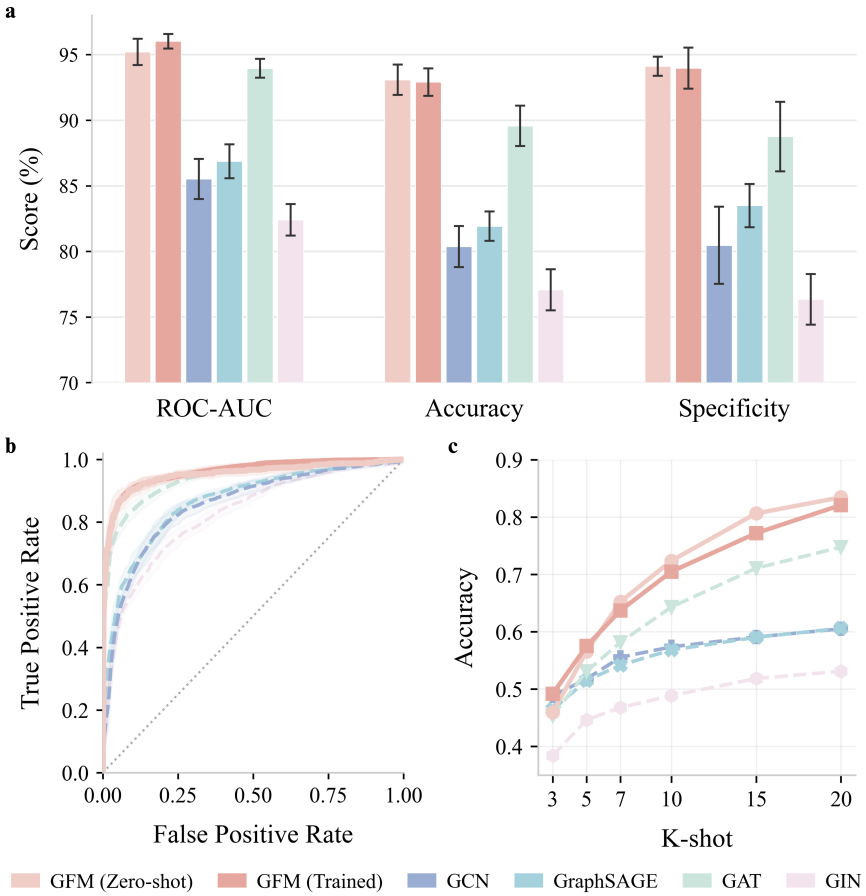

Figure 5. The pretrained GFM generalizes to unseen protein structural fold classes across 144 tissue-specific interaction networks. a, Classification performance at $K = 20$ labeled examples per fold class. Mean ROC-AUC, accuracy, and specificity are shown for all six models, evaluated on test fold classes that were entirely absent during training. Error bars denote one standard deviation across five evaluation seeds. **b, Macro-averaged ROC curves across all test fold classes**. Solid lines denote GFM variants; dashed lines denote supervised baselines. The dotted diagonal denotes random performance. **c, Few-shot generalization curves on unseen fold classes**. Mean accuracy is shown as a function of labeled examples per fold class ($K \in \{3, 5, 7, 10, 15, 20\}$), evaluated on test fold classes using a logistic probe trained on K support examples per class. Solid lines denote GFM variants; dashed lines denote supervised baselines.

In the previous three datasets we tested the generalization across graphs and feature modalities while maintaining at least partial overlap between the functional categories present during training and those evaluated at test time. The Fold-PPI dataset

consists of a stricter evaluation setting in which the fold classes used for testing are disjoint from those available during training. It is a graph meta-learning benchmark introduced by Huang and Zitnik[26] comprising 144 tissue-specific protein–protein interaction networks with protein structural fold class labels drawn from the SCOP database[27]. The benchmark consists of a strictly disjoint class split in which the fold classes available during training are entirely absent from the test set, leaving no overlap between the structural categories the model observes and those on which it is evaluated. This design helps evaluate cross-class structural generalization under a disjoint label split: a model that learns fold-class-specific patterns during training cannot use them to classify test proteins, because test fold classes have no training analog.

The pretrained GFM in zero-shot mode achieved a mean accuracy of 83.4% and a mean ROC-AUC of 95.5% on the held-out test fold classes at $K = 20$ labeled examples per class (Fig. 5a,b). GFM trained reached 82.1% accuracy and 95.8% ROC-AUC under the same protocol. Both values substantially exceed the strongest supervised baseline, GAT[12], which reached 74.8% accuracy and 94.1% ROC-AUC. GCN[10] and GraphSAGE[13] reached 60.5% and 60.6% accuracy respectively, and GIN[11] reached 53.1%. The ROC-AUC gaps between GFM variants and the weaker baselines are particularly large, exceeding 9 percentage points over GCN and GraphSAGE and 13 percentage points over GIN. Macro-averaged ROC curves confirm that both GFM variants maintain separation from all baselines across the full threshold range, with the advantage most pronounced at low false positive rates (Fig. 5b).

The few-shot generalization curve reveals how this advantage builds with label availability (Fig. 5c). At $K = 3$, all models cluster within a narrow accuracy band of approximately 45–50%, reflecting the difficulty of classifying unseen fold classes from very few examples. A clear separation emerges by $K = 5$, where GFM zero-shot reaches 56.6% and GFM trained reaches 57.5%, while GAT reaches 53.1% and GCN reaches 51.8%. This gap widens monotonically as K increases. At $K = 20$, both GFM variants exceed GAT by more than 7 percentage points and exceed GCN, GraphSAGE, and GIN by more than 20 percentage points. The trajectories of GCN and GraphSAGE flatten substantially above $K = 10$, indicating that additional labeled examples provide diminishing returns for models whose embeddings do not organize the test fold class structure effectively. The GFM few-shot curves continue to rise steeply through $K = 20$, consistent with the interpretation that pretrained topology-based representations provide a strong basis for fold class discrimination even under the disjoint label setting.

A notable feature of the Fold-PPI result is that GFM zero-shot accuracy is marginally higher than GFM trained accuracy at $K = 20$ (83.4% versus 82.1%), which reverses the relationship observed in SagePPI, OGBN-Proteins, and StringGO. This inversion is consistent with the disjoint class structure of the benchmark. Supervised fine-tuning on Fold-PPI trains the backbone to encode the 19 training fold classes, which introduces class-specific geometric structure that does not directly transfer to the 5 test fold classes. GFM zero-shot embeddings, derived from graph topology alone, are not biased toward any specific fold class and generalize more uniformly across the class boundary. This observation reinforces the argument that topology-based pretraining produces representations that are broadly transferable precisely because they do not encode dataset-specific label structure.

t-SNE visualization of GFM trained embeddings shows that the 29 training fold classes are organized into compact, well-separated clusters in the embedding space, with one cluster per SCOP fold class label (Supplementary Fig. S11). Each cluster corresponds to a distinct structural class as defined by SCOP secondary structure composition and topology. The clustering emerges from supervised fine-tuning of the backbone on training fold classes and demonstrates that the pretrained model can learn fold-class geometry when label information is available. Baseline models show substantially less compact and separated cluster structure under the same visualization, with GCN and GraphSAGE embeddings producing diffuse distributions that do not resolve individual fold classes clearly.

GFM trained embeddings also capture a tissue-prevalence gradient that is not present in baseline embeddings (Supplementary Fig. S12). The centralization strength, defined as the negative Spearman correlation between fold class tissue prevalence and mean centroid distance, reaches 0.296 for GFM trained, indicating that proteins belonging to fold classes expressed across more tissue contexts are placed closer to their class centroid in embedding space. GAT achieves the highest baseline centralization strength at 0.126, which is less than half the GFM value. GCN achieves 0.027, GraphSAGE 0.075, and GIN 0.096. This gradient reflects the known relationship between structural fold prevalence and network centrality in tissue-specific PPI graphs: proteins with broadly expressed structural architectures tend to occupy topologically central positions across tissue contexts[30], and GFM trained embeddings encode this positional signal more faithfully than embeddings derived from supervised training alone.

Fold-PPI tests two forms of generalization at once. The 5 test fold classes are disjoint from the 19 training fold classes, and the 144 tissue-specific interaction networks were not seen during pretraining. Under these conditions, GFM zero-shot outperforms the best supervised baseline at every value of K without any fold-class-specific training, and the gap over weaker baselines exceeds 20 percentage points at $K = 20$. The tissue prevalence gradient and t-SNE cluster structure confirm that GFM embeddings encode biologically coherent fold-level organization from the interaction topology, consistent with the findings on SagePPI, OGBN-Proteins, and StringGO. Taken together, these results demonstrate that topological pretraining produces representations that generalize to entirely new protein fold categories defined by SCOP structural classification, without any

fold-class-specific adaptation.

Discussion

This work introduces a graph foundation model that can successfully transfer knowledge across graph domains. A single model pretrained on diverse heterogeneous non-biological graphs can be reused on 18 held-out biological graphs and 6 non-biological graphs, spanning 37 downstream evaluation tasks. The central finding is that topology-derived pretraining provides reusable graph representations even when the pretraining graphs contain no proteins, biological interactions, or functional annotations. In zero-shot mode, the frozen pretrained model outperforms supervised GNN baselines trained directly on the target biological graphs across all benchmark graphs. With limited labels, it maintains a consistent few-shot advantage. With supervised fine-tuning, it further improves performance on SagePPI, OGBN-Proteins, StringGO, and all other datasets. These results show that graph topology contains transferable structural information that can support cross-domain graph learning when it is encoded through a shared topological interface.

The mechanism behind successful cross-domain knowledge transfer from non-biological pretraining graphs to held-out biological networks lies in what the structural prompts encode. Degree, clustering coefficient, k-core number, ego-network statistics, PageRank, and community membership are defined purely from graph topology. They carry no reference to what nodes represent or what edges mean. A high-degree node in a citation network and a high-degree node in a protein interaction network occupy topologically analogous positions in their respective graphs, even though one represents a scientific paper and the other represents a hub protein. The contrastive pretraining objective aligns text-derived topological descriptions with graph-derived neighborhood embeddings across nine topologically diverse graphs simultaneously. This forces the backbone to learn representations that are sensitive to topological role rather than to any single graph's semantics. When applied to biological networks, this topology-sensitive representation space preserves properties that have biological meaning because biological function is itself organized topologically. In protein interaction networks, hub proteins often occupy functionally central positions, community structure can reflect shared pathway membership, and diffusion patterns can capture the propagation of functional signal across connected proteins[37]. These biological relationships are not provided during pretraining. Their recovery suggests that the topological signals learned from citation and co-purchase networks overlap with structural signals that carry functional information in protein networks.

The GO hierarchy depth analysis in OGBN-Proteins provides mechanistic evidence that the GFM uses multi-scale topological organization when predicting more specific functional annotations (Fig. 3e). For the GFM trained model, mean ROC-AUC increased from 75.4% on shallow GO terms to 76.7% at medium depth and 80.5% at deep terms, whereas GIN showed no comparable depth-dependent pattern across the same stratification. Shallow GO terms correspond to broad functional categories and are assigned to many proteins, making them more accessible from local neighborhood information. Deep GO terms correspond to more specific functions and are assigned to fewer proteins, making their prediction more dependent on information distributed across longer network paths and community boundaries. The topological prompts are designed to represent these multi-scale signals through diffusion statistics and community indicators. A model trained on a single graph with a fixed label set may not be able to learn transferable multi-scale structure, because its supervision is tied to one graph-specific label distribution. The depth stratification result therefore links the performance gain to a mechanism consistent with the intended topological inductive bias, rather than to an aggregate improvement alone.

The Fold-PPI result reveals a distinctive property of topology-based pretraining. GFM zero-shot outperforms GFM trained at $K = 20$ on Fold-PPI, reversing the general pattern observed in SagePPI, OGBN-proteins, and the Molecular Function and Cellular Component tasks in StringGO (Fig. 5a,c). This inversion is consistent with the disjoint class structure of the benchmark. Supervised fine-tuning on the 19 training fold classes can introduce geometric bias toward those specific structural categories. When evaluation shifts to 5 entirely unseen fold classes, this class-specific bias can reduce transfer to the held-out label space. In contrast, GFM zero-shot embeddings are not optimized toward the training fold classes and therefore preserve a more class-agnostic representation. This observation has practical implications for deployment: when the label vocabulary at test time differs from the label vocabulary available during training, topology-based pretrained representations may provide a more robust initialization than representations obtained by supervised fine-tuning on a fixed label set. This setting is relevant in biology and medicine, where new disease categories, newly annotated gene functions, and newly characterized cell types are frequently introduced after models are trained.

The multifunctionality gradient provides an independent validation of the representation space that does not depend on classification metrics. Across both SagePPI and OGBN-Proteins, GFM zero-shot places highly annotated proteins in denser embedding neighborhoods (Spearman $r = +0.263$ and $r = +0.105$ respectively; Figs. 2b, 3c), while GIN produces negative correlations on both graphs ($r = -0.290$ and $r = -0.020$). This pattern is consistent with the known relationship between multifunctionality and network centrality in protein interaction networks: proteins that participate in many biological processes tend to occupy high-degree, high-betweenness positions that are well connected across functional modules[30,38]. GFM zero-shot captures this centrality signal directly from topology without any supervision from functional labels. Supervised baselines do

not, because their training objective optimizes label prediction rather than structural position. The recurrence of a positive correlation in both SagePPI and OGBN-Proteins, despite differences in graph topology and annotation structure, indicates that this pattern is a stable property of the GFM representation rather than a dataset-specific artifact.

Co-enrichment analysis further indicates that the GFM embedding space preserves established biological relationships even though biological annotations are not used during pretraining. In SagePPI, transcriptional processes cluster together and immune processes form a separate coherent block (Fig. 2c). In OGBN-Proteins, an immune module and a cytokine module show strong mutual enrichment consistent with their known coupling in inflammatory signaling. In StringGO, the cellular component ontology resolves into three spatially distinct modules corresponding to mitochondria, the nuclear and cytoplasmic compartment, and the plasma membrane (Fig. 4f). These are the major spatial compartments of eukaryotic cells. They emerge from protein interaction network topology alone, without any annotation of subcellular localization provided to the model. In the molecular function ontology, transcription factor activity terms cluster according to their shared mechanistic role in RNA polymerase II-driven gene expression (Fig. 4e). None of these biological relationships are encoded in the topological prompts or the pretraining objective. They emerge because proteins that share biological function tend to interact preferentially with each other, and the learned representations preserve this interaction-based proximity faithfully.

The Biological Process ontology in StringGO reveals an instructive boundary condition. GFM zero-shot achieves 80.5% mean ROC-AUC on Biological Process, exceeding all supervised baselines (Supplementary Fig. S7), and the co-enrichment analysis confirms that the zero-shot embedding space organizes Biological Process annotations into biologically coherent clusters (Supplementary Fig. S9). The GFM Trained model, however, does not improve over zero-shot on this ontology. The Biological Process label space is substantially larger and more heterogeneous than Molecular Function or Cellular Component, and the validation signal from its imbalanced annotation density creates more instability during the two-stage supervised fine-tuning procedure used here. This observation implies that the pretrained representations encode the necessary functional organization, and approaches such as label-balanced sampling, ontology-aware curriculum scheduling, or separate adaptation heads for large heterogeneous label spaces are well positioned to recover this gap. The result thus identifies a specific regime where improved adaptation procedures will be most effective, rather than a fundamental limitation of the representations themselves.

The extended evaluation adds 33 downstream task-results across additional biological and non-biological graphs, providing a broader test of where the topology-based representations succeed (Supplementary Figs. S13–S23). The GFM exceeds all four supervised baselines on every metric on the BioGRID human interactome[39] and the IntAct human interactome[40], two protein interaction sources independent of STRING. The same advantage holds on the TRRUST human regulatory network[41] across all three GO ontologies, despite TRRUST representing transcription factor to target relationships rather than physical interactions. The pattern extends across species. Cross-species transfer is preserved on five additional STRING interactomes spanning *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans* (Supplementary Figs. S21, S19, S15, S17, S16), together with the budding yeast *Saccharomyces cerevisiae* (Supplementary Fig. S18). The GFM advantage is consistent across vertebrate, invertebrate, and fungal proteomes despite the pretraining distribution containing no biological graphs. This indicates that the topological signals captured by the structural prompts carry predictive content for functional organization across kingdoms.

A consistent pattern in the extended evaluation concerns the relative behavior of supervised baselines. On every dense human PPI graph evaluated here, including BioGRID, IntAct, and the larger STRING species, the message-passing baselines (GCN, GraphSAGE, and GAT) collapse to near-random mean ROC-AUC values of approximately 50%, while GIN remains the only competitive supervised baseline (Supplementary Figs. S20, S13, S21, S19). This collapse is a known consequence of over-smoothing in deep message-passing networks operating on densely connected graphs[21], where repeated neighborhood aggregation drives node representations toward a common mean. The GFM is less susceptible to this failure mode because its representation is anchored to topology-derived prompts that encode community structure, k-core position, and diffusion statistics directly rather than relying entirely on message passing to recover them. The over-smoothing failure of three of the four baselines on alternative PPI sources strengthens the mechanistic interpretation: GFM performance on dense biological graphs is not driven by a stronger GraphSAGE backbone alone but by the explicit topological signal that the prompts provide.

The non-protein benchmarks tested in the extended evaluation extend the GFM advantage beyond the protein interaction setting. On the DeezerEurope and LastFMAsia musical preference networks[42] and on the TRRUST binary transcription factor classification task, the GFM exceeds all four supervised baselines on accuracy and macro F1. On PubMed[43], FacebookPagePage[44], and the GitHub developer network[44], the GFM matches the strongest supervised baseline within 2% on every metric. These three latter graphs fall within the citation, co-authorship, and social network families that constitute the pretraining distribution, so the GFM is competing with supervised baselines that train directly on the same graph type it was pretrained on. The pretrained representations therefore remain competitive even when the baselines hold an evaluation advantage. Taken together, the extended evaluation indicates that topology-based pretraining produces a graph foundation model that successfully transfers knowledge across protein, regulatory, social, and citation network types without degradation,

achieving the largest absolute gains where the supervised baselines lack the prior structure needed to recover the underlying organization.

The topological prompt framework is most appropriate for graphs in which edges encode informative relational structure, including interaction networks, citation graphs, and co-authorship graphs. When edges are sparse, noisy, or weakly related to the downstream task, the community detection and diffusion statistics used in the prompts may become less stable. In such settings, adapter tuning may benefit from noise-robust community detection, uncertainty-aware diffusion features, or edge reliability weighting. The current framework also assumes access to the full graph topology during adaptation. Extending topological prompting to streaming, partially observed, or privacy-constrained graphs, where the complete adjacency structure is unavailable at inference time, remains an important direction for future work.

The implications of this work extend beyond the benchmarks evaluated here. In biology and medicine, labeled data are often limited because experimental annotation, clinical curation, and cohort generation are costly. A model pretrained once on publicly available non-biological graphs and transferred to biological networks with minimal adaptation can reduce dependence on large task-specific training sets. The few-shot results support this point directly: the pretrained backbone provides useful representations with as few as one labeled example per class. This is particularly relevant for emerging biological settings where functional annotations remain incomplete, including newly sequenced organisms, newly characterized cell types in single-cell atlases[45], and disease-specific interaction networks for which curated labels are limited. The structural prompt framework is not restricted to protein interaction networks. The same class of topology-derived prompts can be computed for other graph-structured biological systems, including gene regulatory networks, cell–cell communication graphs, drug–target interaction networks[46], and clinical knowledge graphs. Whether representations learned from citation and co-purchase networks transfer to these additional biological and clinical graph types will require direct evaluation.

Outlook

We have introduced a graph foundation model that learns transferable topological representations from heterogeneous non-biological graphs and reuses them on held-out biological networks with minimal adaptation. The model encodes each node's topological role through prompts that summarize local connectivity, community membership, and diffusion-based position, and aligns these descriptions with graph-derived neighborhood embeddings through contrastive pretraining. Across four biological network benchmarks spanning human protein interaction networks, a species-stratified STRING-derived benchmark, a multi-ontology STRING network with amino acid composition features, and 144 tissue-specific interaction networks with disjoint structural fold labels, the frozen pretrained backbone outperforms supervised GNN baselines trained directly on the target graphs in zero-shot evaluation. In few-shot settings, the pretrained embeddings provide useful representations from as few as one labeled example per class, while supervised baselines require substantially more labeled examples to approach comparable performance. The learned embedding space preserves established biological relationships, including transcriptional modules, immune and cytokine signaling clusters, mitochondrial compartment organization, and cellular localization structure, without using functional annotations during pretraining. In ogbn-proteins, performance increased with GO hierarchy depth, indicating that the topology-derived representations retain multi-scale structural information relevant to fine-grained functional prediction. In Fold-PPI, the zero-shot model achieved higher accuracy than the fine-tuned GFM on entirely unseen fold classes, while maintaining comparable ROC-AUC, showing that topology-based pretraining can preserve transfer across disjoint label spaces. These results establish topological graph properties as a transferable inductive structure for reusable graph models that generalize across biological and non-biological domains.

Methods

Overview

The Graph Foundation Model (GFM) is designed to learn node representations that transfer across graphs with incompatible node feature spaces. The central challenge is that standard graph neural networks encode node features directly, which means the learned parameters are specific to the feature schema of the training graph and cannot be reused on a graph with a different feature type or dimensionality. The GFM addresses this by replacing raw node features with a feature-agnostic topological description of each node's position and role in the graph. This description is derived entirely from topology and is expressible as a natural language string that can be encoded by a language model into a fixed-dimensional vector, regardless of the original feature space. The model is pretrained once by aligning these topology-derived text embeddings with graph-structure embeddings produced by a message-passing backbone across nine heterogeneous graphs simultaneously. After pretraining, the backbone is frozen and reused on unseen graphs by fitting a lightweight per-graph adapter that maps the new graph's node features into the pretrained representation space. This adapter is tuned without labels using a contrastive objective on the target graph's topology. The resulting embeddings can then be used directly with a linear probe for zero-shot evaluation or fine-tuned with a task-specific head when labeled data is available.

Topological prompt construction

For each node i in a graph $G = (V, E)$, a structured natural-language profile is constructed from two categories of topological descriptors: local context and global context.

Local context descriptors capture the immediate topological environment of the node. These include the node degree d_i, the local clustering coefficient c_i, the k-core number[31], first- and second-order ego-network statistics (vertex count, edge count, and density at radius 1 and 2), and the PageRank score p_i computed with damping factor 0.85 over 40 power iterations[47].

Global context descriptors capture the node's position within the large-scale community structure of the graph. Two independent community detection algorithms are applied. Label Propagation[33] is run for 20 iterations to assign each node to a community, and the community identity, size, and internal edge density are recorded. SCoDA, a streaming community detection algorithm[32], is applied independently and its community identity, size, and density are recorded separately. Using two algorithms reduces dependence on any single community detection method and provides complementary views of mesoscale structure.

These descriptors are assembled into a natural-language node profile of the form:

```
Node profile:  local(deg=12, cc=0.167, core=4, ego1V=13, ego1E=22, ego1D=0.046,
ego2V=214, ego2E=4103, ego2D=0.181, pr=0.00084); global(lp_comm=5, lp_size=312,
lp_dens=0.021; scoda_comm=8, scoda_size=189, scoda_dens=0.018); graph(N=2708, E=5429,
avgd=4.01, trans=0.241, q25=2, q50=3, q75=5, spec_gap=1.23).
```

Each profile is tokenized using the all-MiniLM-L6-v2 Sentence Transformer[48], producing a 384-dimensional embedding z_i^{text} for each node. The full computational definitions of each descriptor are provided in Supplementary Methods.

Graph Feature Embedding

For a graph with N nodes and raw node feature matrix $X \in \mathbb{R}^{N\times d}$, a per-graph adapter $A : \mathbb{R}^{N\times(d+384)} \to \mathbb{R}^{N\times 1024}$ is applied to the concatenation of raw features and topological tokens:

$$\tilde{X} = A\big([X \,\|\, Z^{\text{text}}]\big) \tag{1}$$

where $Z^{\text{text}} \in \mathbb{R}^{N\times 384}$ is the matrix of MiniLM token embeddings and $\|$ denotes row-wise concatenation. The adapter is a single linear layer with output dimension 1024. A separate adapter is instantiated for each graph during pretraining. For held-out graphs, a new adapter is initialized from the mean of the pretrained adapter weights and tuned without labels on the target graph's topology.

The adapted features $\tilde{X}$ are processed by a two-layer GraphSAGE backbone[13] with hidden dimension 512 and output dimension 256, using dropout rate 0.6 at each layer:

$$g = \text{Normalize}\big(\text{GraphSAGE}(\tilde{X}, E)\big) \in \mathbb{R}^{N\times 256} \tag{2}$$

A text projection head maps the same adapted features to a matching space:

$$z = \text{Normalize}\big(W_{\text{proj}}\tilde{X}\big) \in \mathbb{R}^{N\times 256} \tag{3}$$

where $W_{\text{proj}} \in \mathbb{R}^{1024\times 256}$. The final node embedding is a weighted concatenation of the graph stream and the text stream:

$$e_i = [\alpha \cdot g_i \,\|\, (1-\alpha)\cdot z_i] \in \mathbb{R}^{512} \tag{4}$$

with $\alpha = 0.7$.

Contrastive pretraining objective

The model is pretrained by aligning the graph structure embedding g_i with the text stream embedding z_j of structurally related nodes. Positive pairs (i, j) are identified using Personalized PageRank (PPR) with restart probability 0.15, run for 100 iterations per anchor node. The top-96 highest-PPR nodes relative to each anchor are treated as positives[49]. This selection captures multi-hop topological proximity without relying on direct adjacency.

The contrastive loss is an InfoNCE objective[50,51] with temperature $\tau = 0.1$:

$$\mathscr{L}_{\text{NCE}} = -\frac{1}{2}\left(\log\frac{\exp(g_i\cdot z_j/\tau)}{\sum_k \exp(g_i\cdot z_k/\tau)} + \log\frac{\exp(z_j\cdot g_i/\tau)}{\sum_k \exp(z_j\cdot g_k/\tau)}\right) \tag{5}$$

averaged symmetrically over both directions. For graphs with more than 20,000 nodes, a sampled variant with 1024 in-batch negatives is used to remain memory-efficient. For smaller graphs the full node bank is used as the negative set, identical to the original InfoNCE formulation.

A Laplacian smoothing regularizer penalizes large differences between embeddings of adjacent nodes:

$$\mathscr{L}_{\text{smooth}} = \lambda \cdot \frac{1}{|E|}\sum_{(u,v)\in E} \|g_u - g_v\|^2 \tag{6}$$

with $\lambda = 5 \times 10^{-3}$. The total training loss is $\mathscr{L} = \mathscr{L}_{\text{NCE}} + \mathscr{L}_{\text{smooth}}$.

Pretraining protocol

The model is pretrained on nine heterogeneous graphs spanning academic citation networks (Cora[43], CiteSeer[43], DBLP[43], ogbn-arxiv[23]), scientific co-authorship networks (CoauthorCS[52], CoauthorPhysics[52]), e-commerce product co-purchase networks (AmazonComputers[52], AmazonPhoto[52]), and a Wikipedia computer science hyperlink network (WikiCS[53]). None of these graphs contain proteins, biological interactions, or functional annotations.

At each training step, one graph is sampled uniformly at random. A batch of 1024 anchor nodes is drawn from that graph, PPR positives are computed, and the loss is evaluated on the anchor-positive pairs. Training proceeds for 250 epochs with 128 steps per epoch. The Adam optimizer[54] is used with learning rate 10^{-5} and weight decay 5×10^{-4}. Mixed-precision training[55] is applied throughout. The checkpoint with the lowest training loss is retained for downstream evaluation.

Downstream adaptation

For each evaluation graph, a new linear adapter is initialized from the mean of the pretrained adapter weights and tuned without labels using the same contrastive objective applied to the target graph's topology. For SagePPI the adapter is tuned for 2000 steps. For OGBN-Proteins, StringGO, and Fold-PPI the adapter is tuned for 5000 steps. The backbone and text projection head remain frozen during adapter tuning.

Zero-shot evaluation uses the frozen backbone with the adapted embeddings and a logistic regression linear probe fitted on the training split. Supervised fine-tuning proceeds in two stages. In the first stage, the backbone and text projection are frozen and only the adapter and task-specific classification head are trained. In the second stage, all parameters are unfrozen and trained jointly with a lower learning rate for the backbone (10^{-5}) and text projection (10^{-5}) than for the adapter (10^{-4}) and head (10^{-3}). Both stages use binary cross-entropy loss and are monitored on the validation split by mean ROC-AUC. The best validation checkpoint is retained for test evaluation.

Few-shot evaluation trains a logistic regression probe on K labeled examples per class drawn from the training split, with $K \in \{1,5,10,20\}$ for SagePPI and OGBN-Proteins, $K \in \{1,5,10,20\}$ for StringGO, and $K \in \{3,5,7,10,15,20\}$ for Fold-PPI. For each value of K, the probe is evaluated on the full test split.

Baseline implementations

Four supervised GNN baselines are evaluated on each dataset: Graph Convolutional Networks (GCN)[10], Graph Isomorphism Networks (GIN)[11], Graph Attention Networks (GAT)[12], and GraphSAGE[13]. All four baselines use a two-layer architecture with hidden dimension 256. Each baseline is trained from scratch on the target graph using raw node features only, without topological prompts or pretrained weights. Training uses the Adam optimizer with learning rate 5×10^{-3}, weight decay 5×10^{-4}, and dropout rate 0.5. A maximum of 500 training epochs is applied with early stopping based on validation mean ROC-AUC. GAT uses 4 attention heads in the first layer and 1 head in the second. The same hyperparameters are applied across all four evaluation datasets without dataset-specific tuning.

Evaluation metrics

The primary evaluation metric is mean ROC-AUC, computed as the unweighted average of per-label ROC-AUC scores across all valid labels (labels with at least two represented classes in the test set). Classification accuracy is reported as a secondary metric using per-label decision thresholds tuned on the validation set by maximizing binary F1 score. For Fold-PPI, accuracy is the primary metric consistent with the original benchmark protocol. All experiments use random seed 42.

Zero-shot lightweight evaluation (ZS-LE) protocol

Supplementary Figures S13 through S23 report GFM performance under the zero-shot lightweight evaluation (ZS-LE) protocol. ZS-LE uses the pretrained GFM as a frozen feature extractor without updating any GFM parameter at adaptation or evaluation time. A small two-layer multilayer perceptron classifier with hidden dimension 256, ReLU activation, dropout rate 0.3, and a layer-normalized input is fitted on the frozen GFM embeddings of the training split, using binary cross-entropy for multilabel tasks and cross-entropy for single-label tasks. The GFM backbone, text projection head, and tuned adapter remain frozen throughout. This protocol corresponds to the zero-shot evaluation convention used in genomics foundation models including Geneformer[56], scGPT[57], CellFM[58], and the Geneformer v2 scaling study[59], where the foundation model is held frozen and a lightweight classifier head provides the only task-specific parameters.The ZS-LE protocol is reported in the supplementary because it represents the standard zero-shot evaluation in the broader foundation-model literature, while the linear probe protocol used in the main paper is more conservative and isolates the linear separability of the embedding space.

Evaluation datasets

SagePPI

The SagePPI dataset is derived from the protein–protein interaction network introduced alongside the GraphSAGE framework[13]. It comprises 24 tissue-specific human PPI graphs with a total of 56,944 nodes and 818,716 edges, averaging approximately 2,372 proteins per graph. The standard split allocates 20 graphs to training, 2 to validation, and 2 to testing. Node features are 50-dimensional vectors composed of positional gene sets, motif gene sets, and immunological signatures. Multilabel classification targets are 121 binary labels derived from the MSigDB Gene Ontology biological process collection[29].

OGBN-Proteins

The OGBN-Proteins benchmark is from the Open Graph Benchmark[23]. The underlying network is derived from the STRING database version 11[24]. It contains 132,534 protein nodes across eight species connected by 39,561,252 undirected edges representing biologically meaningful associations including physical interactions and genetic co-expression. The original dataset contains no node features. Node representations are constructed by aggregating the 8-dimensional edge features of each node's incident edges. Multilabel classification targets are 112 binary GO function labels. The benchmark uses a species-stratified split in which test proteins are drawn from species absent during training. No modifications were made to the standard split.

StringGO

The StringGO dataset is constructed from the human interactome in the STRING database[24]. Nodes represent approximately 18,000 human proteins. Edges represent predicted and experimental functional associations, yielding approximately 8,000,000 interactions. Node features are 20-dimensional amino acid composition vectors. Labels are derived from the Gene Ontology resource[25] and evaluated separately across three ontologies: Molecular Function, Biological Process, and Cellular Component. Each ontology defines its own train/validation/test split over proteins.

Fold-PPI

The Fold-PPI dataset was introduced with the G-Meta graph meta-learning framework[26]. It comprises 144 tissue-specific human protein–protein interaction networks. Nodes represent proteins and edges represent physical interactions localized to a specific tissue context. Node features are 512-dimensional conjoint triad vectors[60] encoding amino acid composition and local sequence triplet statistics. Labels correspond to 30 protein structural fold classes from the SCOP database[27]. The benchmark uses a strictly disjoint class split in which 19 fold classes are available during training and 5 entirely distinct fold classes are reserved for evaluation. No fold class identity is shared between the training and test sets.

Extended cross-source human PPI networks

The BioGRID human interactome[39] provides an independent measurement of physical and genetic interactions in *Homo sapiens* curated from primary literature, with 17,474 nodes and 839,636 undirected edges after restriction to UniProt-resolvable human proteins. The IntAct human interactome[40] provides a third independent measurement compiled by the European Bioinformatics Institute, with 18,056 nodes and 570,214 undirected edges after the same UniProt resolution. For both sources, node features are 20-dimensional amino acid composition vectors and labels are GO annotations from the Gene Ontology resource[25], filtered separately by ontology (Molecular Function, Biological Process, Cellular Component) using the same minimum-count thresholds as StringGO.

Cross-species STRING networks

Six additional species-specific STRING v12.0 interactomes[24] are evaluated as separate single-species graphs: *Mus musculus* (taxon 10090, 16,580 proteins), *Rattus norvegicus* (taxon 10116, 7,866 proteins), *Danio rerio* (taxon 7955, 3,156 proteins), *Drosophila melanogaster* (taxon 7227, 3,731 proteins), *Caenorhabditis elegans* (taxon 6239, 4,156 proteins), and *Saccharomyces cerevisiae* (taxon 4932, 6,088 proteins). For each species, STRING edges above the combined score threshold of 500 are retained and proteins are resolved to canonical UniProt accessions through STRING aliases and UniProt secondary accession mapping. Node features are 20-dimensional amino acid composition vectors and labels are species-specific GO annotations from the Gene Ontology Annotation database[61], filtered separately by ontology.

Regulatory network (TRRUST)

The TRRUST human transcription factor target network[41] is a manually curated regulatory network in which edges represent transcription factor to target gene regulatory relationships rather than physical interactions. After symbol to UniProt resolution, the graph contains 2,804 nodes and 8,188 undirected edges. Two complementary classification tasks are defined on this graph. The first uses GO annotations as labels (TRRUSTHuman_GO) with the same Molecular Function, Biological Process, and Cellular Component ontologies and minimum-count thresholds as the other biological networks. The second is a binary classification task (TRRUSTHuman_TF) in which each node is labeled as a transcription factor or a regulated target according to its role in the source TRRUST table.

Social and citation networks

Six non-protein single-label classification benchmarks are evaluated to test the boundary of the GFM advantage. PubMed[43] is a citation network of diabetes-related papers with 500-dimensional TF-IDF features and three topic labels. The Facebook Page-Page network and the GitHub developer network[44] are hyperlink and follower graphs respectively, both with multi-hot bag-of-words features over webpage or biographical content. DeezerEurope and LastFMAsia[42] are music streaming service user-user follower networks with multi-hot features over artists liked by each user, with binary or 18-class genre labels. These five datasets are obtained from the Stanford Network Analysis Project repository[62] in their MUSAE-format release.

Data availability

All datasets analyzed in this study are publicly available from established repositories. The SagePPI protein–protein interaction network was obtained from the GraphSAGE framework repository introduced by Hamilton *et al.*[13]. The OGBN-Proteins benchmark was obtained from the Open Graph Benchmark[23]. The StringGO dataset was constructed from the STRING database version 11[24] with Gene Ontology annotations from the Gene Ontology Consortium[25]. The Fold-PPI dataset was obtained from the G-Meta repository introduced by Huang and Zitnik[26]. All nine pretraining graphs (Cora, CiteSeer, DBLP, ogbn-arxiv, CoauthorCS, CoauthorPhysics, AmazonComputers, AmazonPhoto, and WikiCS) are publicly available through the PyTorch Geometric library[63] and the Open Graph Benchmark[23].

All processed structural token caches, pretrained model checkpoints, and intermediate embedding files required to reproduce the analyses reported in this study are provided within the associated Code Ocean capsule prepared for peer review. Detailed preprocessing procedures and model training protocols are described in the Methods. No new human or animal data were generated for this study.

Code availability

All code used to perform the analyses in this study has been deposited in a Code Ocean capsule and shared with the editors and reviewers as part of the peer-review process. The capsule contains the complete implementation of the Graph Foundation Model, including topological prompt construction, multi-graph pretraining, per-graph adapter tuning, downstream evaluation pipelines, and all scripts used to generate the figures reported in this study.

Upon acceptance of the manuscript, the Code Ocean capsule will be made publicly available with a persistent digital object identifier (DOI) to enable full reproducibility of the results.

References

1. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
2. Yan, R., Islam, M. T. & Xing, L. Deep representation learning of protein-protein interaction networks for enhanced pattern discovery. *Sci. Adv.* **10**, eadq4324, DOI: 10.1126/sciadv.adq4324 (2024).
3. Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
4. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
5. Nicholson, D. N. & Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **18**, 1414–1428 (2020).
6. Brown, T. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).
7. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
8. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
9. Wu, Z. *et al.* A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks Learn. Syst.* **32**, 4–24 (2021).
10. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations* (2017).
11. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (2019).

12. Veličković, P. *et al.* Graph attention networks. In *International Conference on Learning Representations* (2018).

13. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

14. Veličković, P. *et al.* Deep graph infomax. In *International Conference on Learning Representations* (2019).

15. Zhu, Y. *et al.* Deep graph contrastive representation learning. In *ICML Workshop on Graph Representation Learning and Beyond* (2020).

16. Thakoor, S. *et al.* Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations* (2022).

17. Zhu, Z., Lin, K., Jain, A. K. & Zhou, J. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888* (2020).

18. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478* (2021).

19. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. *AI Open* **1**, 57–81 (2020).

20. Wu, Z. *et al.* A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks Learn. Syst.* **32**, 4–24 (2021).

21. Li, Q., Han, Z. & Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

22. Newman, M. E. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).

23. Hu, W. *et al.* Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).

24. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

25. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

26. Huang, K. & Zitnik, M. Graph meta learning via local subgraphs. In *Advances in Neural Information Processing Systems*, vol. 33 (2020).

27. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).

28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

29. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

30. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

31. Batagelj, V. & Zaversnik, M. An O(m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049* (2003).

32. Hollocou, A., Maudet, J., Bonald, T. & Lelarge, M. A linear streaming algorithm for community detection in very large networks. *arXiv preprint arXiv:1703.02955* (2017).

33. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. vol. 76, 036106 (APS, 2007).

34. Dinarello, C. A. Proinflammatory cytokines. *Chest* **118**, 503–508 (2000).

35. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45–54 (2019).

36. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).

37. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).

38. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLOS Comput. Biol.* **3**, e59 (2007).

39. Stark, C. *et al.* Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).

40. Orchard, S. *et al.* The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).

41. Han, H. *et al.* Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

42. Rozemberczki, B. & Sarkar, R. Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, 1325–1334 (2020).

43. Yang, Z., Cohen, W. & Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning* (2016).

44. Rozemberczki, B., Allen, C. & Sarkar, R. Multi-scale attributed node embedding. *J. Complex Networks* **9**, cnab014 (2021).

45. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

46. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).

47. Page, L., Brin, S., Motwani, R. & Winograd, T. The pagerank citation ranking: Bringing order to the web. Tech. Rep., Stanford infolab (1999).

48. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019).

49. Gasteiger, J., Bojchevski, A. & Günnemann, S. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations* (2019).

50. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

51. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (2020).

52. Shchur, O., Mumme, M., Bojchevski, A. & Günnemann, S. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868* (2018).

53. Mernyei, P. & Cangea, C. Wiki-CS: A Wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901* (2020).

54. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations* (2015).

55. Micikevicius, P. *et al.* Mixed precision training. In *International Conference on Learning Representations* (2018).

56. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

57. Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nat. methods* **21**, 1470–1480 (2024).

58. Zeng, Y. *et al.* Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat. Commun.* **16**, 4679 (2025).

59. Chen, H. *et al.* Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology. *Nat. Comput. Sci.* 1–14 (2026).

60. Shen, J. *et al.* Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci.* **104**, 4337–4341 (2007).

61. Huntley, R. P. *et al.* The goa database: gene ontology annotation updates for 2015. *Nucleic Acids Res.* **43**, D1057–D1063 (2015).

62. Leskovec, J. & Krevl, A. Snap datasets: Stanford large network dataset collection. http://snap.stanford.edu/data (2014).

63. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

Supporting Information for

Language-Encoded Structural Topology Enables Generalizable Foundation Models for Graph-Structured Data

Sakib Mostafa, James Zou, Ash A. Alizadeh, Maximilian Diehn, Lei Xing, and Md Tauhidul Islam

Corresponding Author: Md. Tauhidul Islam and Lei Xing.
E-mail: tauhid@stanford.edu and lei@stanford.edu

This PDF file includes:

Figs. S1 to S23

Supplementary Materials

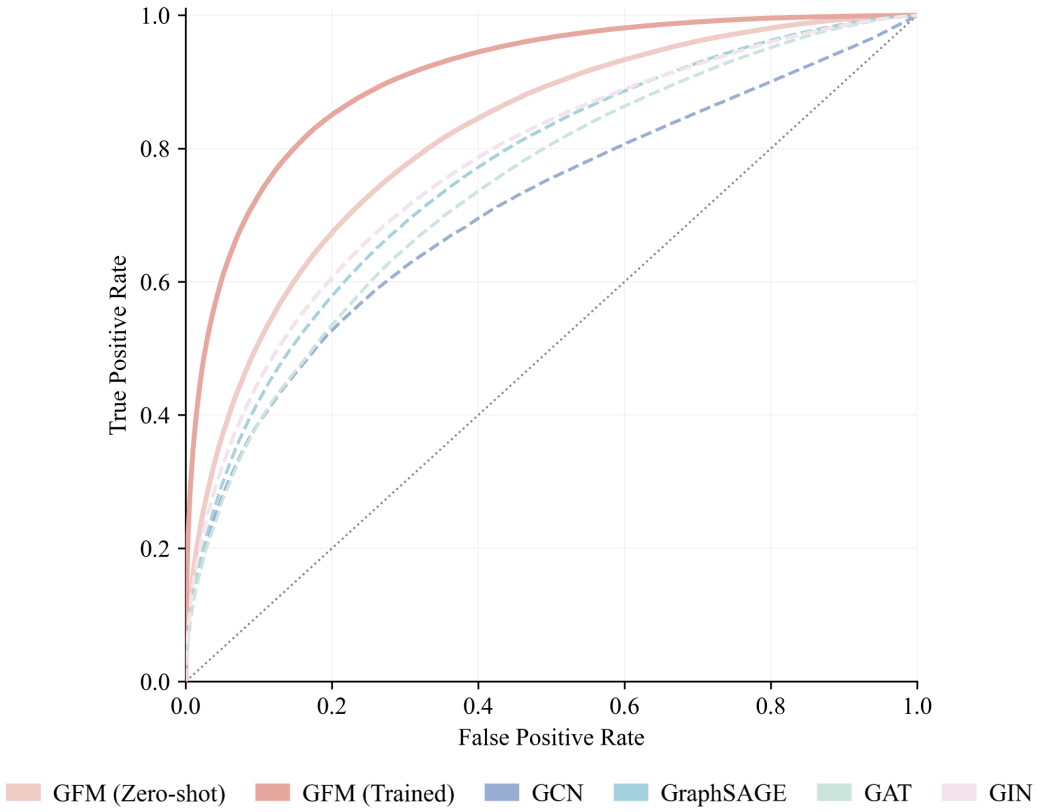

Fig. S1. Full macro-averaged ROC curves for SagePPI. Macro-averaged receiver operating characteristic curves across all 121 GO biological process labels for GFM zero-shot, GFM trained, GCN, GraphSAGE, GAT, and GIN. Curves are computed by interpolating per-label ROC curves onto a common false positive rate grid and averaging. The clear separation between GFM variants (solid lines) and all supervised baselines (dashed lines) is maintained across the full operating range, with the largest absolute gap occurring at low false positive rates where classifier precision requirements are most stringent.

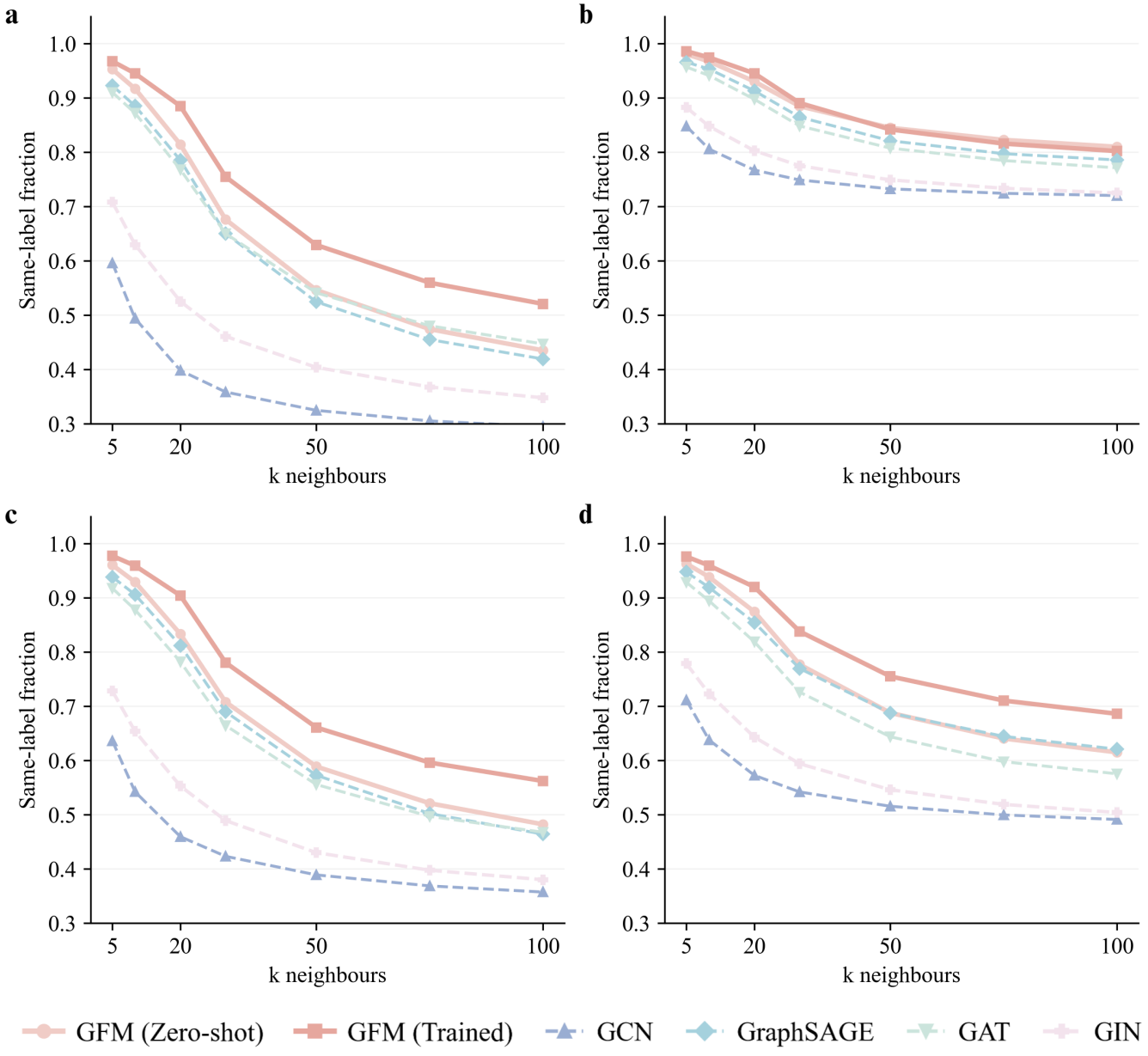

Fig. S2. Same-label neighbor enrichment curves for four GO biological processes in SagePPI. For each model, the same-label fraction is computed as the proportion of k nearest neighbors (in cosine embedding space) that share the same GO process annotation as the query protein, averaged over all proteins annotated with that process. Curves are shown for k ∈ {5, 10, 20, 30, 50, 75, 100}. **a**, DNA repair (GO:0006281). **b**, Cell cycle regulation (GO:0051726). **c**, Nervous system development (GO:0007399). **d**, Negative regulation of transcription (GO:0000122). Solid lines denote GFM variants; dashed lines denote supervised baselines. GFM zero-shot and GFM trained consistently maintain higher same-label fractions at all values of k across all four processes.

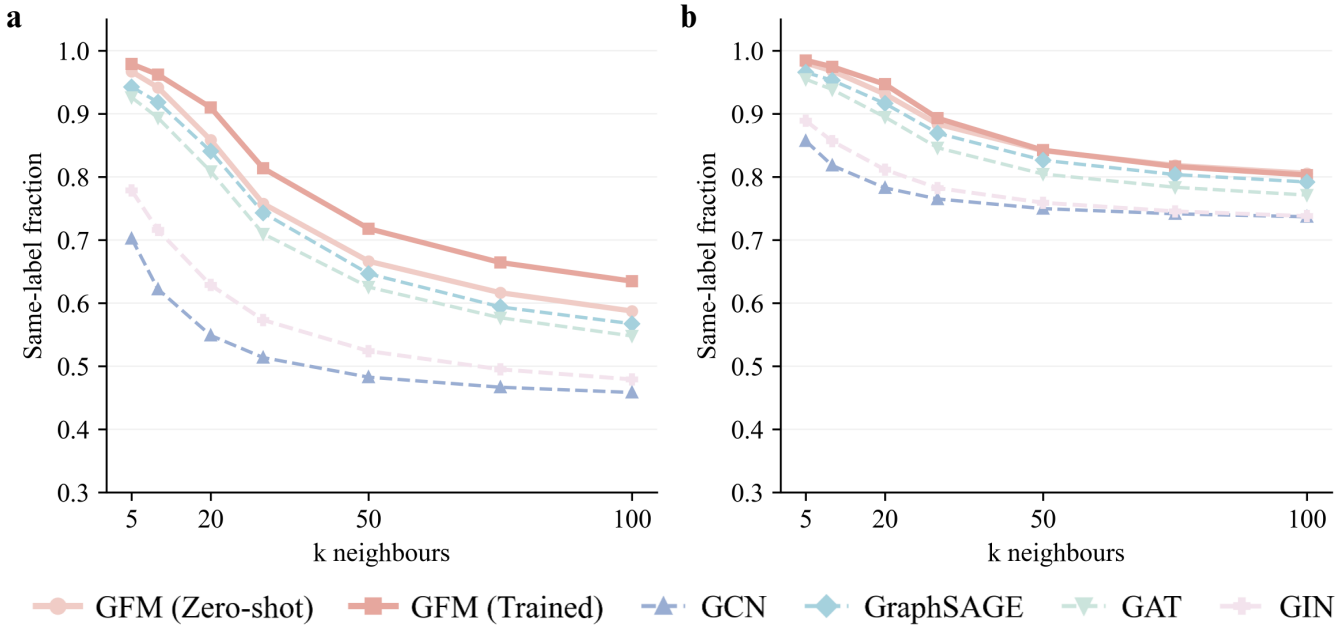

Fig. S3. Same-label neighbor enrichment curves for two additional GO biological processes in SagePPI. a, RNA polymerase II transcription (GO:0006366). **b**, Cell proliferation regulation (GO:0042127). Format and evaluation protocol are identical to Supplementary Fig. 2. The GFM advantage is maintained across both processes, including cell proliferation regulation, which represents a functionally distinct class from the transcriptional and repair processes shown in Supplementary Fig. 2.

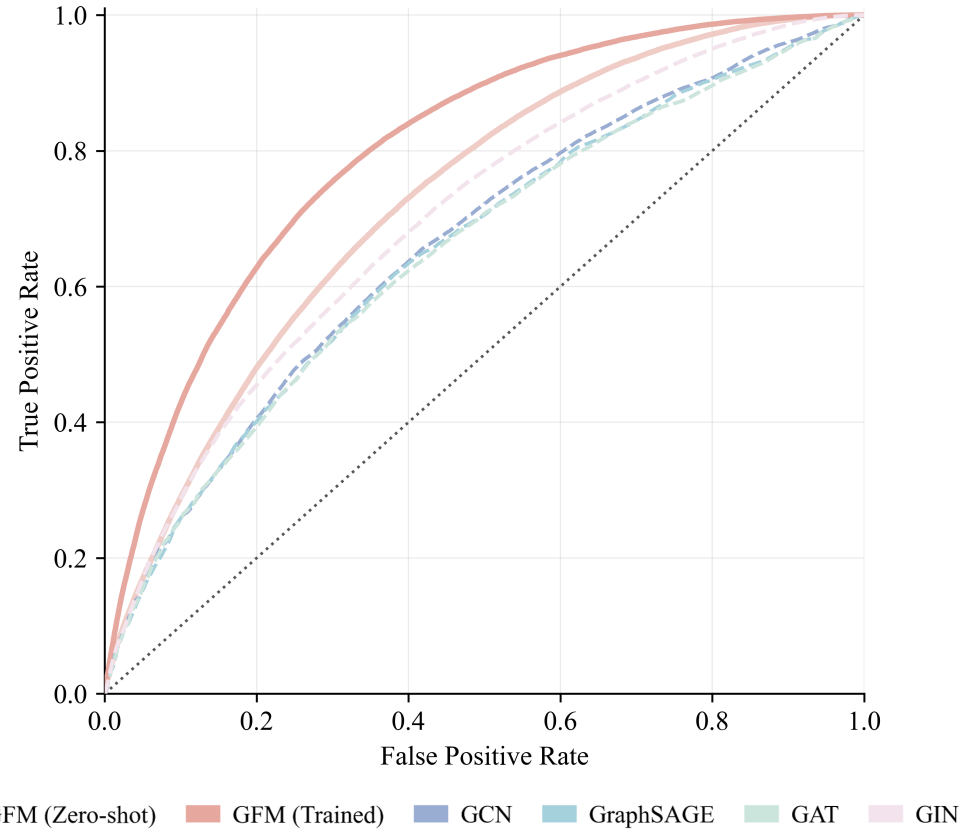

Fig. S4. Full macro-averaged ROC curves for OGBN-Proteins. Macro-averaged receiver operating characteristic curves across all GO biological process labels for GFM zero-shot, GFM trained, GCN, GraphSAGE, GAT, and GIN. The clear separation between GFM variants (solid lines) and all supervised baselines (dashed lines) persists across the full operating range. The gap is widest at low false positive rates, consistent with results on SagePPI (Supplementary Fig. S1).

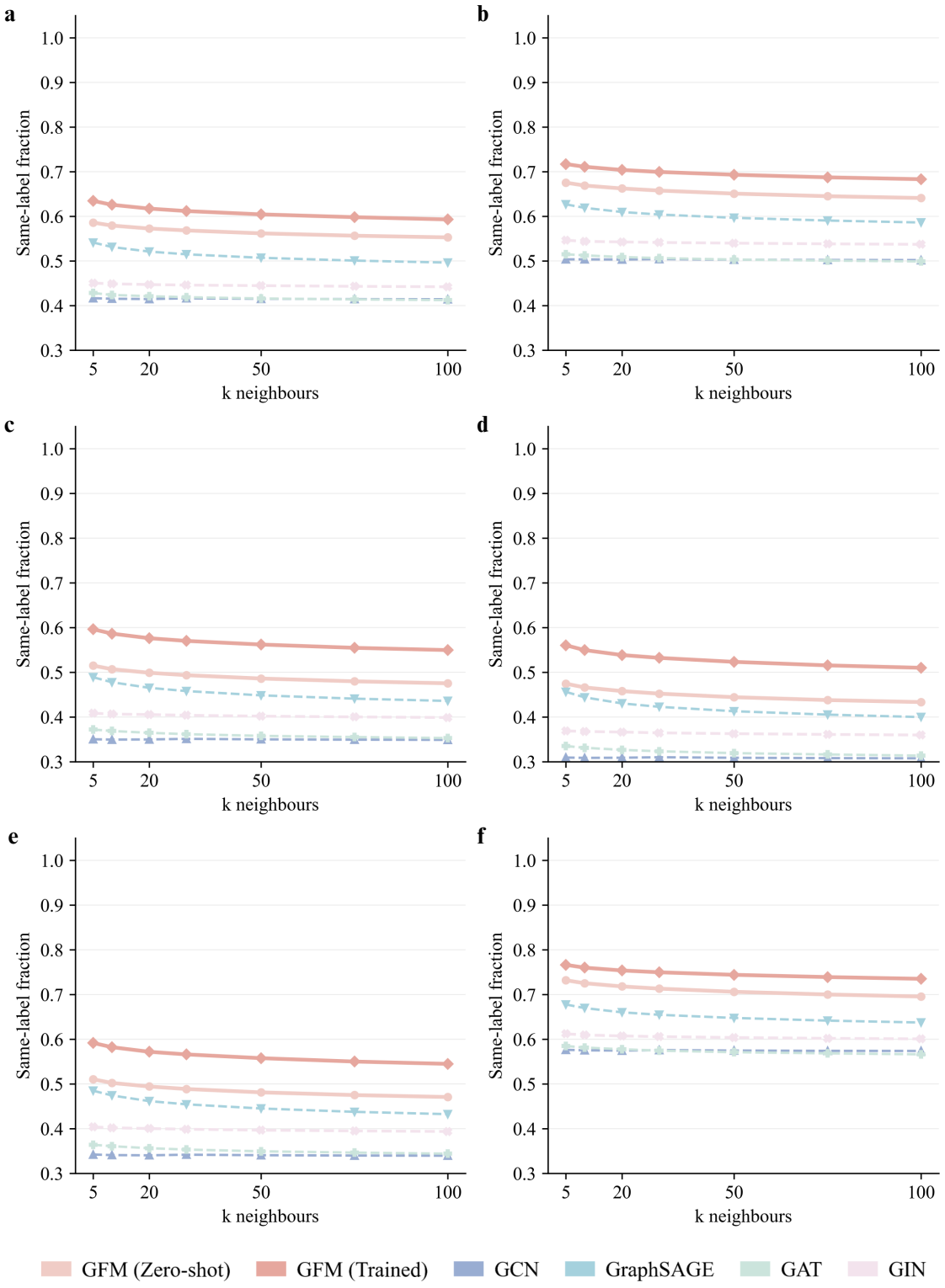

Fig. S5. Same-label neighbor enrichment curves for six GO annotation categories in OGBN-Proteins. For each model, the same-label fraction is computed as the proportion of k nearest neighbors (in cosine embedding space) that share the same GO annotation as the query protein, averaged over all annotated proteins. Curves are shown for $k \in \{5, 10, 20, 30, 50, 75, 100\}$. **a**, Molecular function. **b**, Cellular process. **c**, Binding. **d**, Organic substance metabolic process (GO:0071704). **e**, Cellular metabolic process (GO:0044237). **f**, Nitrogen compound metabolic process (GO:0006807). Solid lines denote GFM variants; dashed lines denote supervised baselines. GFM zero-shot and GFM trained consistently maintain higher same-label fractions at all values of k across all six annotation categories.

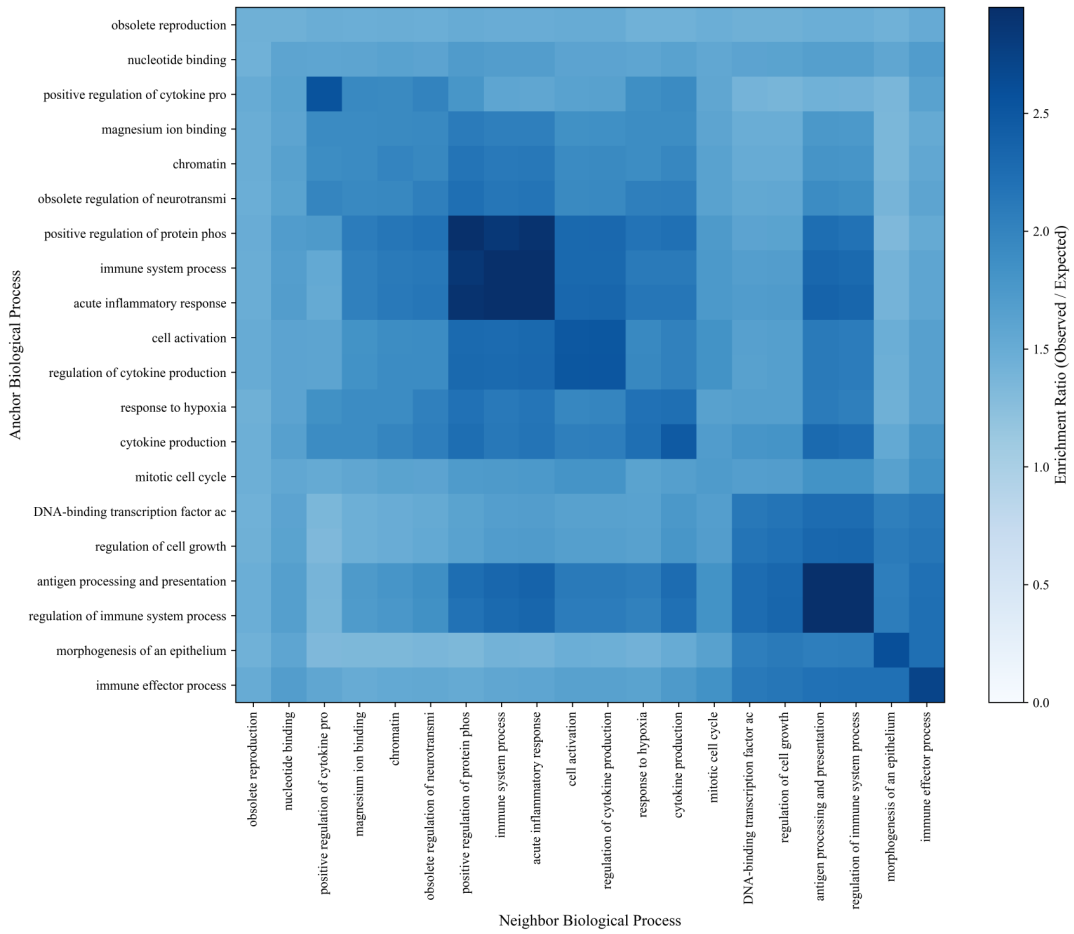

Fig. S6. Cross-functional co-enrichment in GFM trained embeddings for OGBN-Proteins. Each cell reports the enrichment ratio (observed over expected) for the mean fraction of GO-annotated neighbors that a protein from one anchor biological process accumulates within a 10-nearest-neighbor embedding neighborhood annotated by another process. Rows and columns are ordered by hierarchical clustering of the symmetrized co-enrichment matrix. Enrichment ratios substantially above 1.0 indicate that two processes share embedding-space neighborhoods more than expected by chance. The immune cluster—comprising regulation of immune system process, antigen processing and presentation, acute inflammatory response, immune system process, and immune effector process—forms a prominent high-enrichment block. A separate cytokine cluster—positive regulation of cytokine production, regulation of cytokine production, cytokine production, and cell activation—shows strong internal enrichment and cross-enrichment with the immune block, consistent with the known coupling between cytokine signaling and acute inflammatory programs.

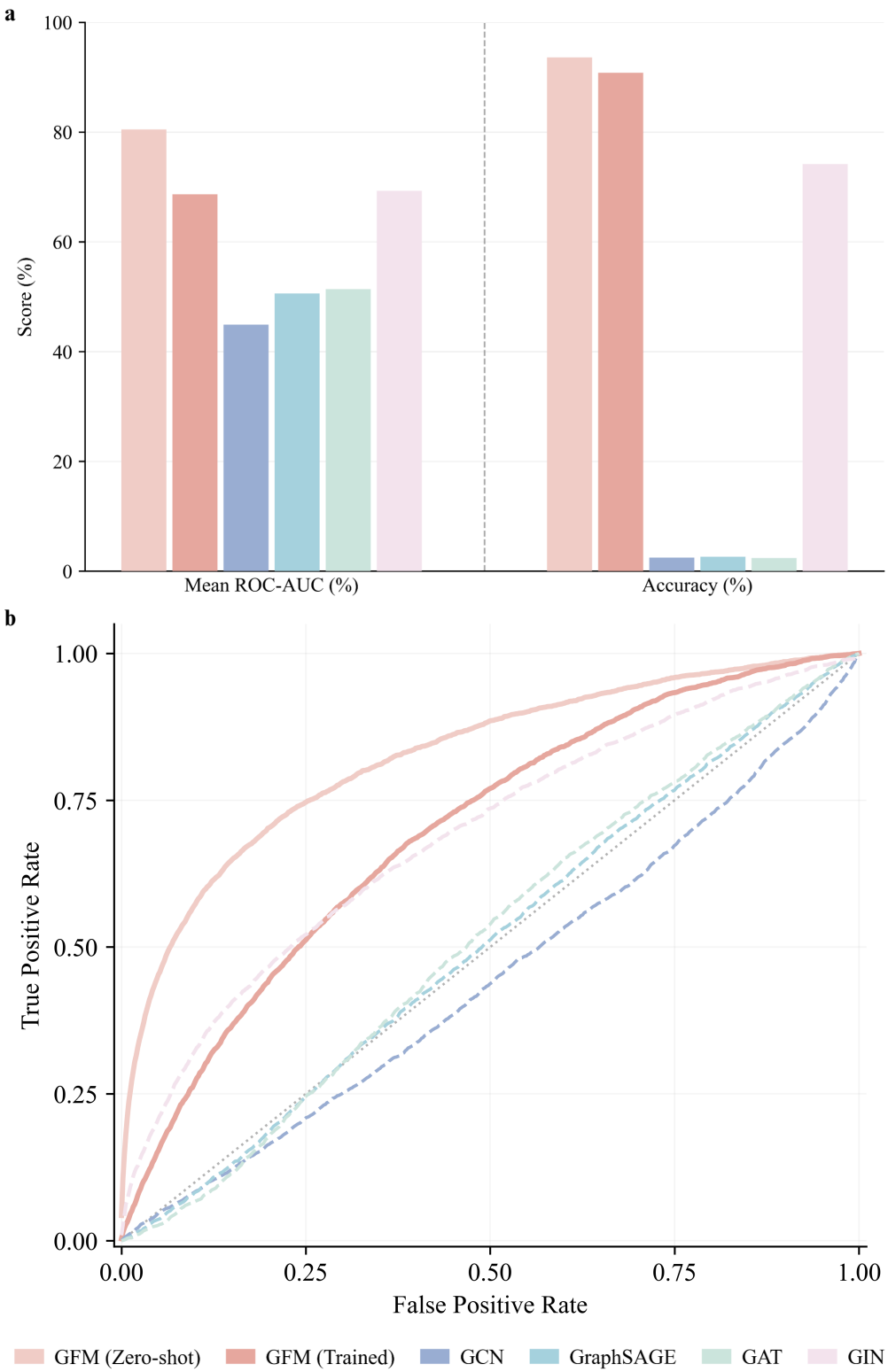

Fig. S7. GFM performance on Biological Process (BP) GO term prediction in StringGO. a, Mean ROC-AUC and accuracy for BP GO term prediction across all six models. GFM zero-shot achieves the highest mean ROC-AUC (80.5%), exceeding all supervised baselines including GIN (69.3%). The GFM trained model does not improve over the zero-shot baseline on this ontology, in contrast to the MF and CC results. This behavior is specific to BP and is discussed in the main text. **b**, Macro-averaged ROC curves for BP. Solid lines denote GFM variants; dashed lines denote supervised baselines. The dotted diagonal denotes random performance. GFM zero-shot maintains a clear separation from all baselines across the full threshold range, with GFM trained remaining between zero-shot and the strongest baseline.

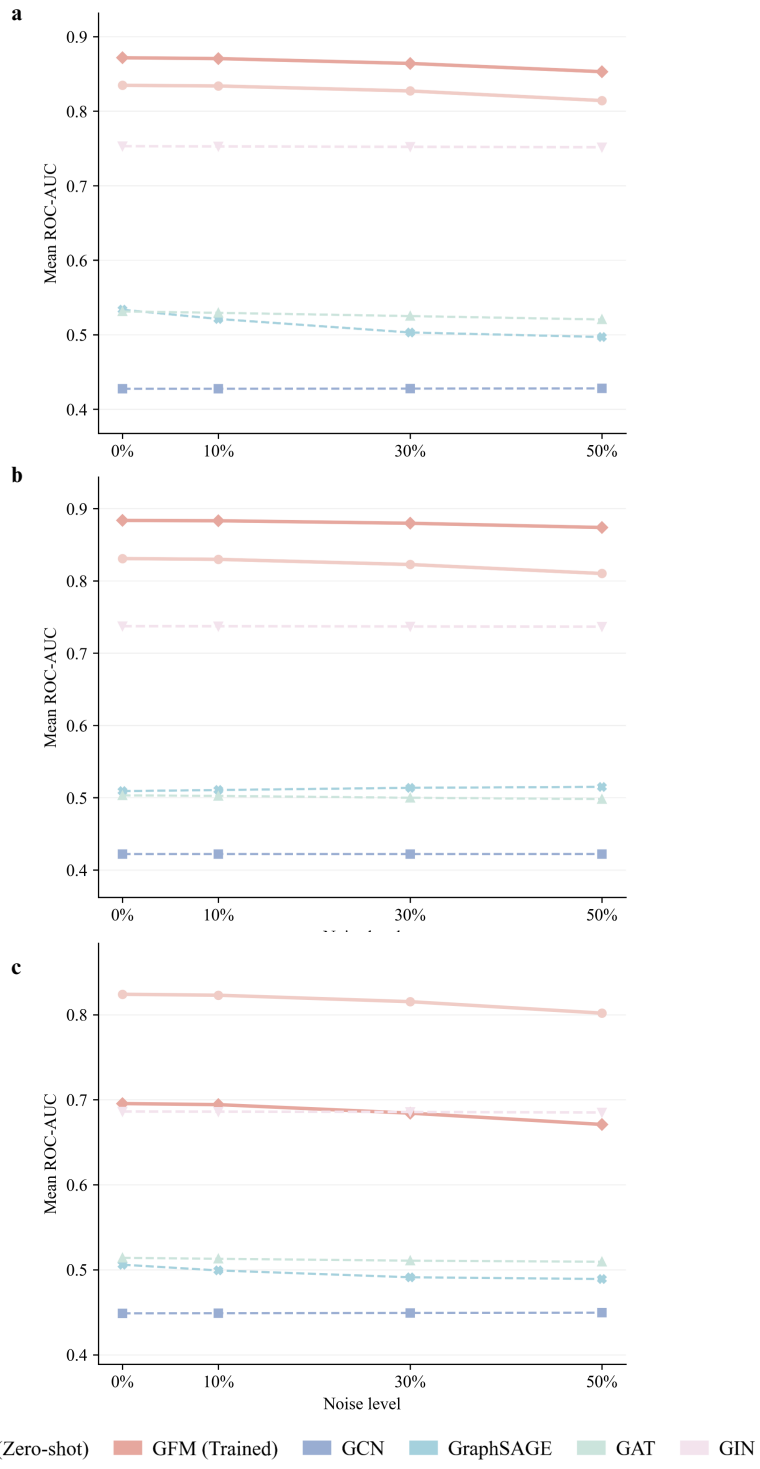

Fig. S8. Noise robustness of GFM and baseline embeddings in StringGO. Mean ROC-AUC is shown as a function of Gaussian noise level applied to the amino acid composition features (x_{raw}), for MF (**a**), CC (**b**), and BP (**c**). Noise is parameterized as a fraction of the per-feature standard deviation across proteins (0%, 10%, 30%, 50%). GFM zero-shot retains a mean ROC-AUC above 0.79 for MF and CC at 50% noise, indicating that its embeddings are largely driven by topology-derived features that are not affected by amino acid feature perturbation. GFM trained similarly maintains strong performance under all noise levels for MF and CC. On BP, both GFM variants show modest decline but remain above all baselines at every noise level. Baseline performance declines more steeply because amino acid composition constitutes their full input.

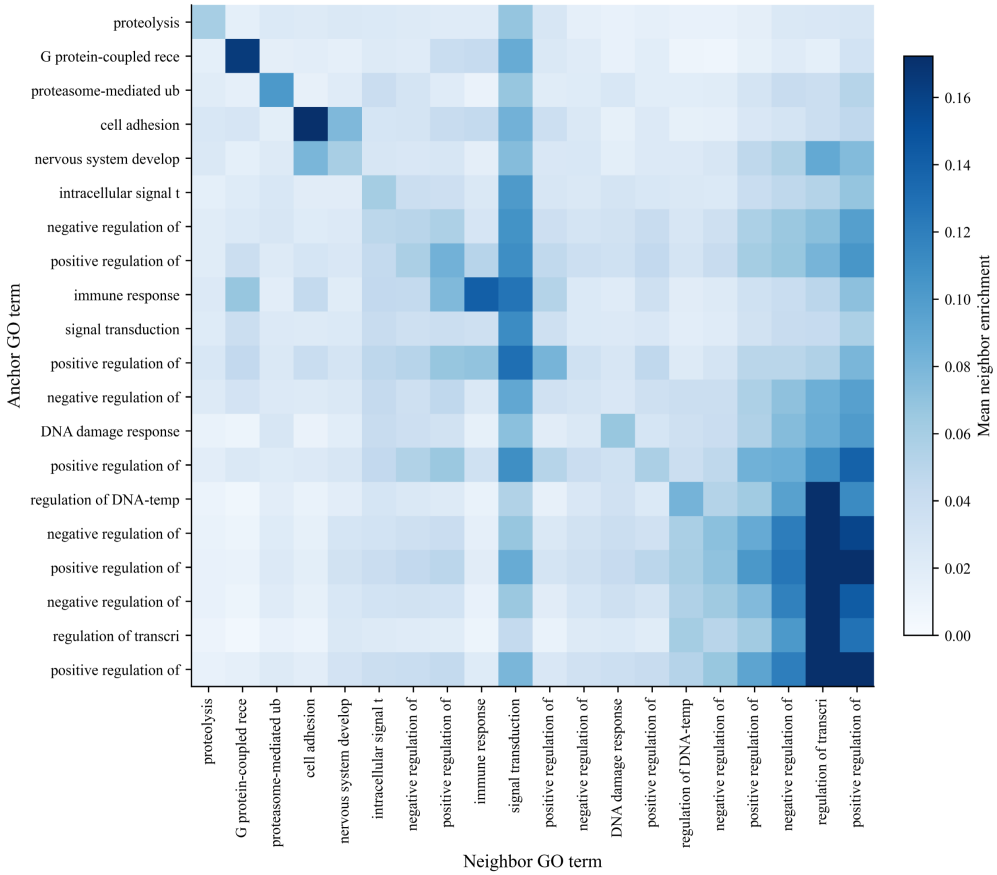

Fig. S9. Cross-functional co-enrichment in GFM trained embeddings for BP in StringGO. Same protocol as Fig. ??e–f. The BP co-enrichment matrix reveals distinct clusters including an immune response module, a signal transduction module, and a transcriptional regulation module. Consistent with the BP performance results, the co-enrichment structure in the GFM zero-shot embedding space is well organized even though the supervised head does not improve on it under the current fine-tuning procedure.

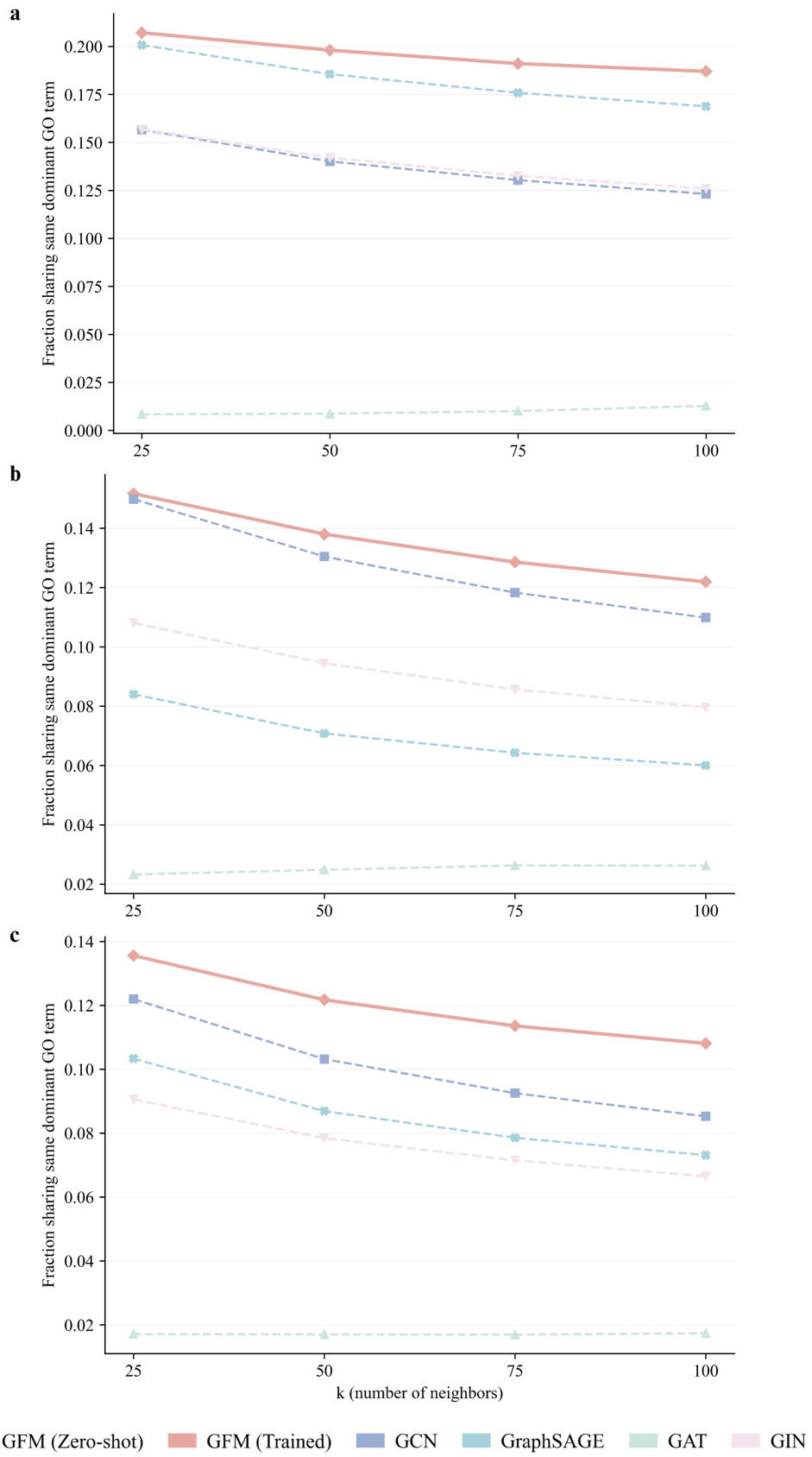

Fig. S10. Cross-ontology neighborhood sensitivity in StringGO embeddings. For each model and ontology, the fraction of k nearest neighbors sharing the same dominant GO term is shown for $k \in \{25, 50, 75, 100\}$ across MF (**a**), BP (**b**), and CC (**c**). The dominant GO term for each protein is defined as its most frequent annotation within the respective ontology. GFM trained consistently places proteins with matching dominant annotations in closer embedding neighborhoods than all supervised baselines across all three ontologies and all values of k.

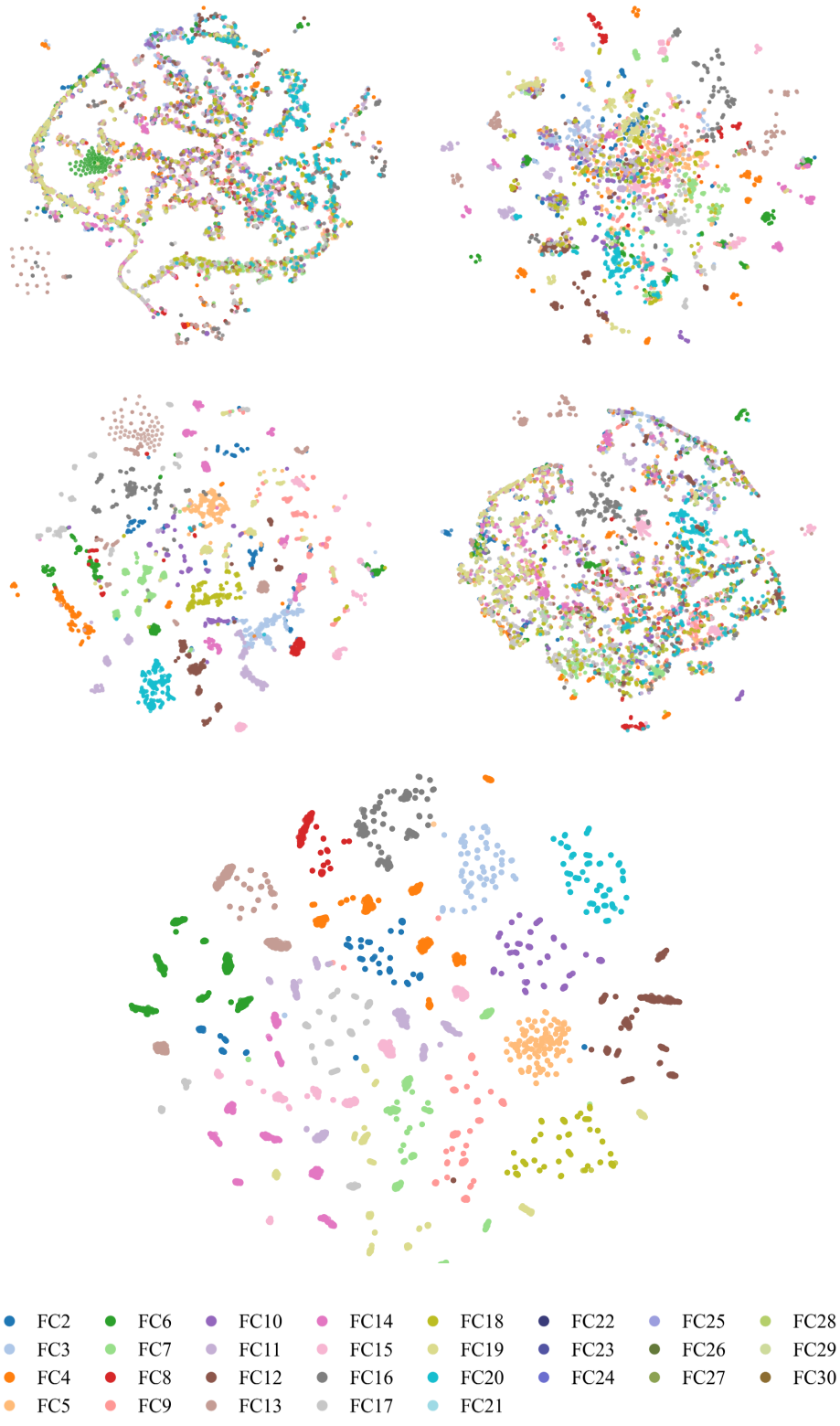

Fig. S11. t-SNE visualization of protein embeddings across SCOP fold classes in Fold-PPI. Each point represents a protein colored by its SCOP fold class assignment (FC2–FC30). The four smaller panels (top two rows) show embeddings from GFM zero-shot, GCN, GraphSAGE, and GAT. The large bottom panel shows GFM trained embeddings, in which fold class clusters are visibly separated and compact. Each cluster corresponds to one of the 29 labeled SCOP fold classes present in the training split. The degree of intra-class compactness and inter-class separation in GFM trained embeddings is substantially greater than in any baseline, indicating that topological pretraining and fine-tuning produce a representation space that organizes protein fold identity from interaction network topology alone.

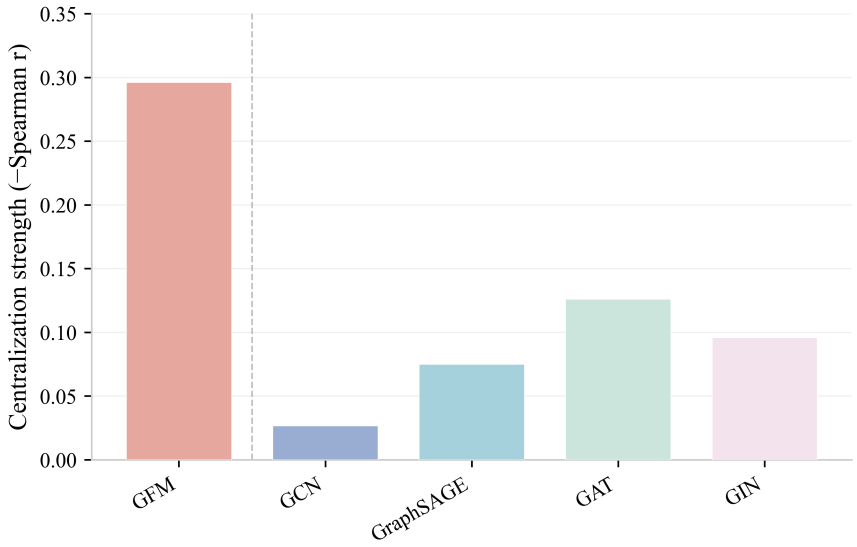

Fig. S12. Tissue-prevalence centralization of fold class embeddings in Fold-PPI. Bars show centralization strength, defined as the negative Spearman rank correlation between the tissue prevalence of each protein's fold class and its mean distance to the fold class centroid in embedding space. A positive value indicates that proteins belonging to fold classes that appear across more tissue contexts are placed closer to their class centroid, reflecting greater embedding compactness for broadly expressed structural folds. GFM trained produces a centralization strength of 0.296, substantially exceeding all supervised baselines. GAT achieves the highest baseline value of 0.126. The dashed vertical line separates the GFM trained model from supervised baselines.

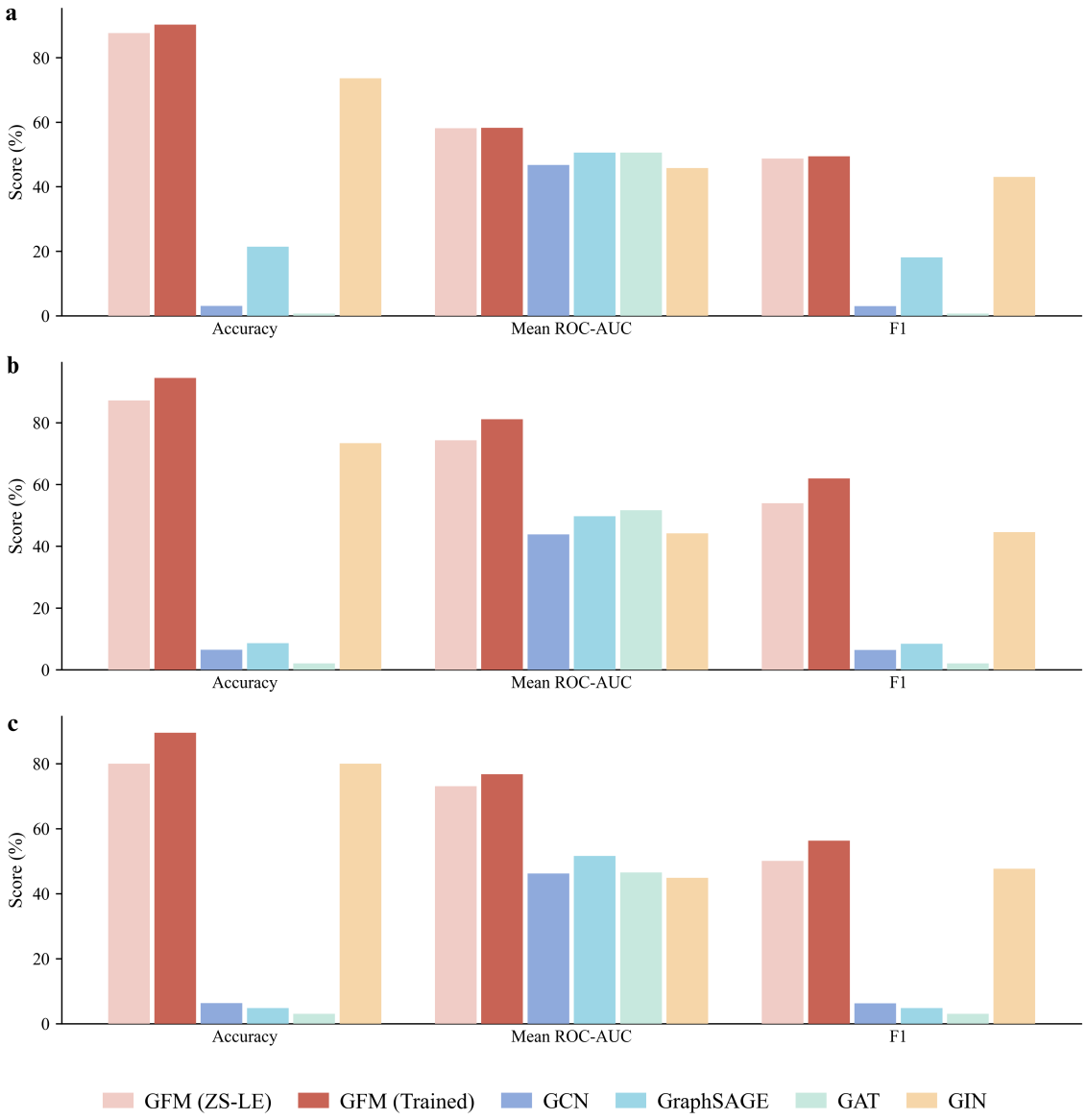

Fig. S13. GFM performance on the IntAct human protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for GFM (ZS-LE), GFM (Trained), GCN, GraphSAGE, GAT, and GIN. ZS-LE denotes a multilayer perceptron classifier trained on frozen GFM embeddings without updating GFM parameters and follows the zero-shot lightweight evaluation protocol described in Supplementary Methods. **a**, Molecular Function (MF) GO term prediction. **b**, Biological Process (BP) GO term prediction. **c**, Cellular Component (CC) GO term prediction. GFM (ZS-LE) and GFM (Trained) exceed all four supervised baselines on all three metrics across every ontology. The four message-passing baselines collapse to near-random accuracy and macro F1 on this dense interaction network, consistent with the over-smoothing behavior documented previously on related human PPI graphs.

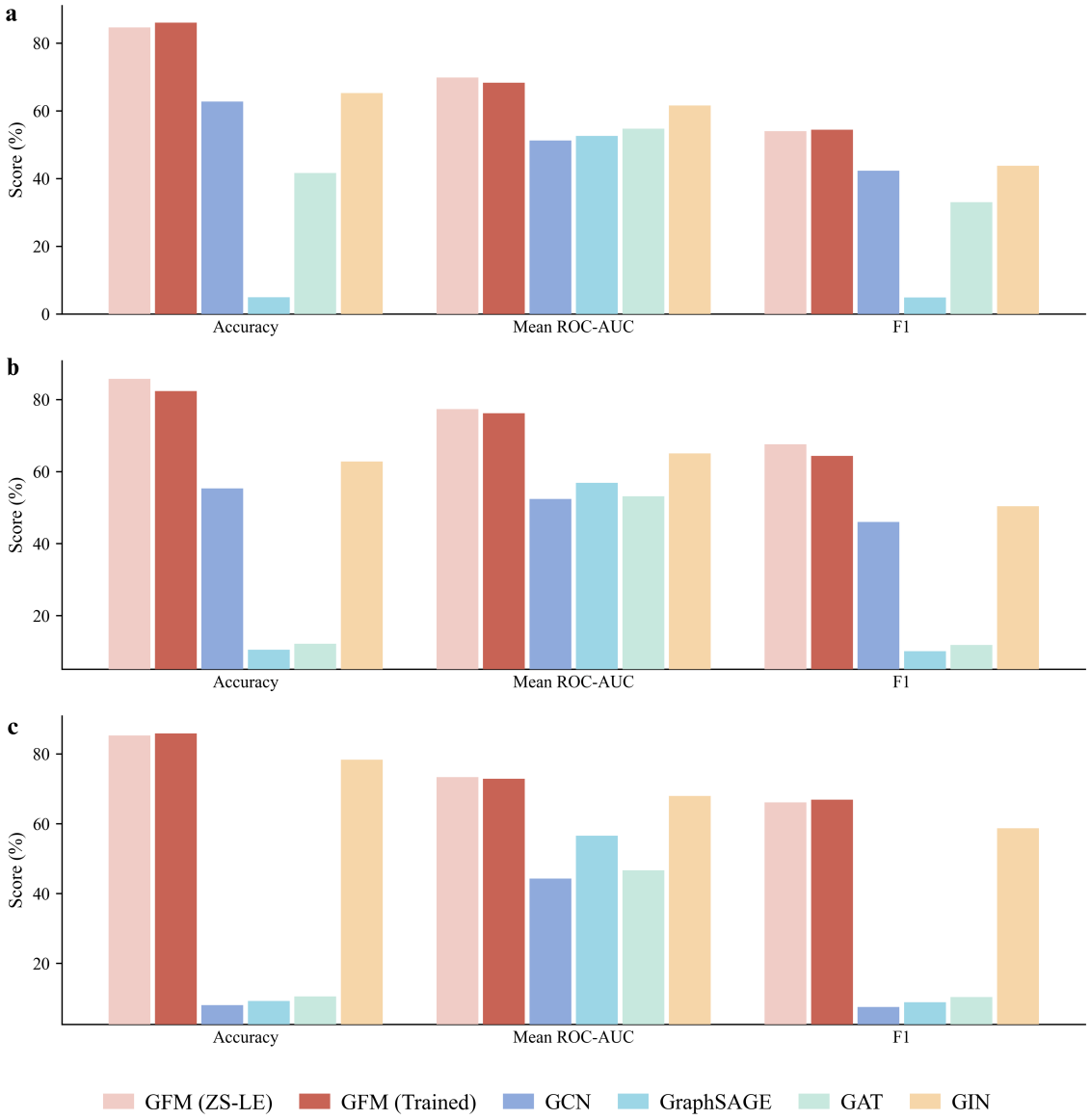

Fig. S14. GFM performance on the TRRUST human regulatory network with GO labels across three ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the human transcription factor target regulatory network with GO annotations as labels. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed every supervised baseline on all three ontologies, with the largest margins on accuracy and macro F1 where GIN reaches only 65% to 80% while both GFM variants exceed 85%. The TRRUST network is sparser and more tree-like than the protein interaction graphs evaluated elsewhere, yet the GFM advantage is preserved across all three ontologies.

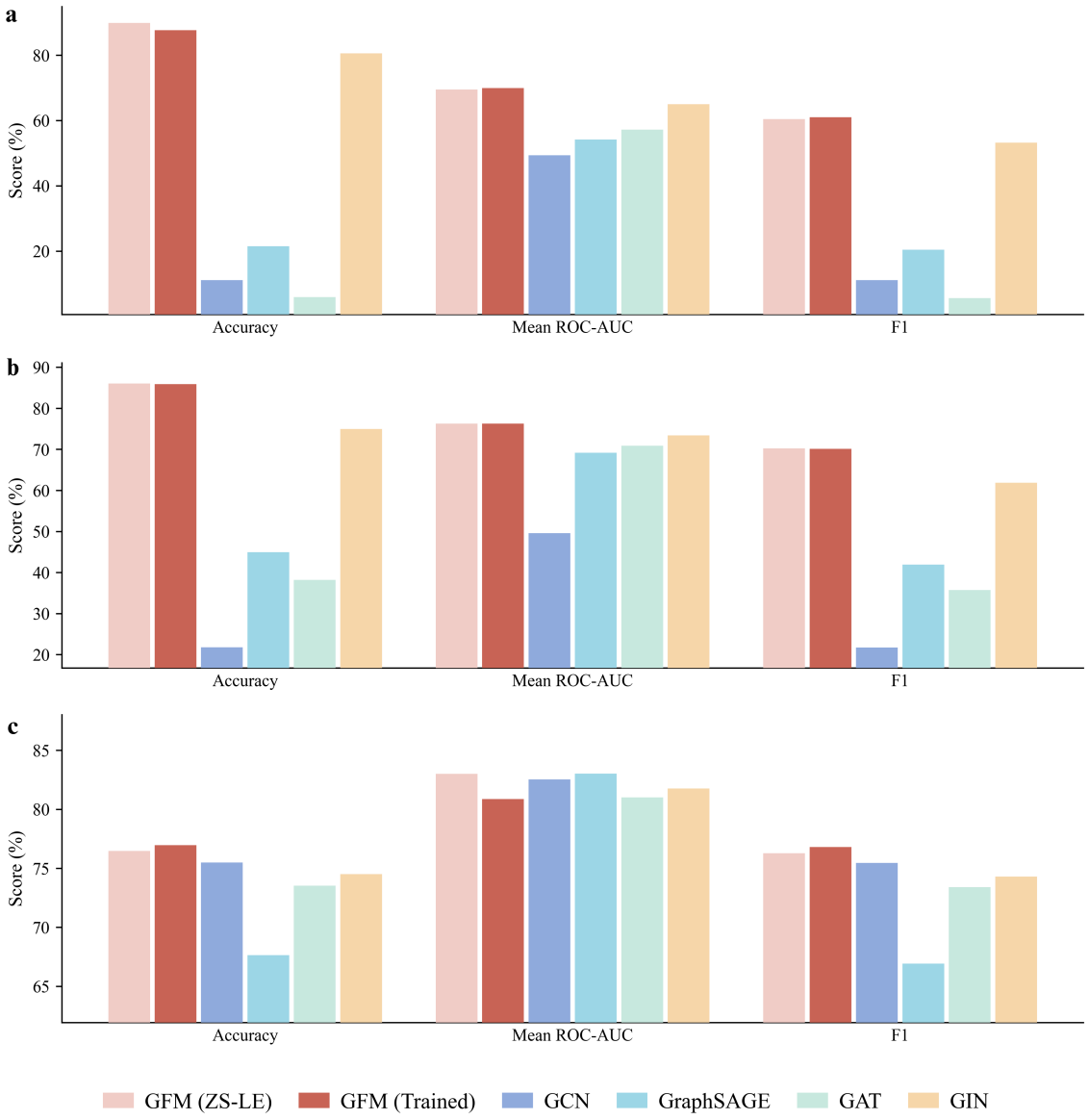

Fig. S15. GFM performance on the zebrafish STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the zebrafish (*Danio rerio*, taxon 7955) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed all supervised baselines on accuracy and macro F1 across all three ontologies. GIN is the strongest baseline on this network but trails the GFM variants by 6% to 14% on accuracy. The GFM advantage persists despite the zebrafish proteome being substantially smaller than the human proteome (3,156 versus approximately 18,000 proteins), indicating that the pretrained representations transfer to vertebrate species independent of network size.

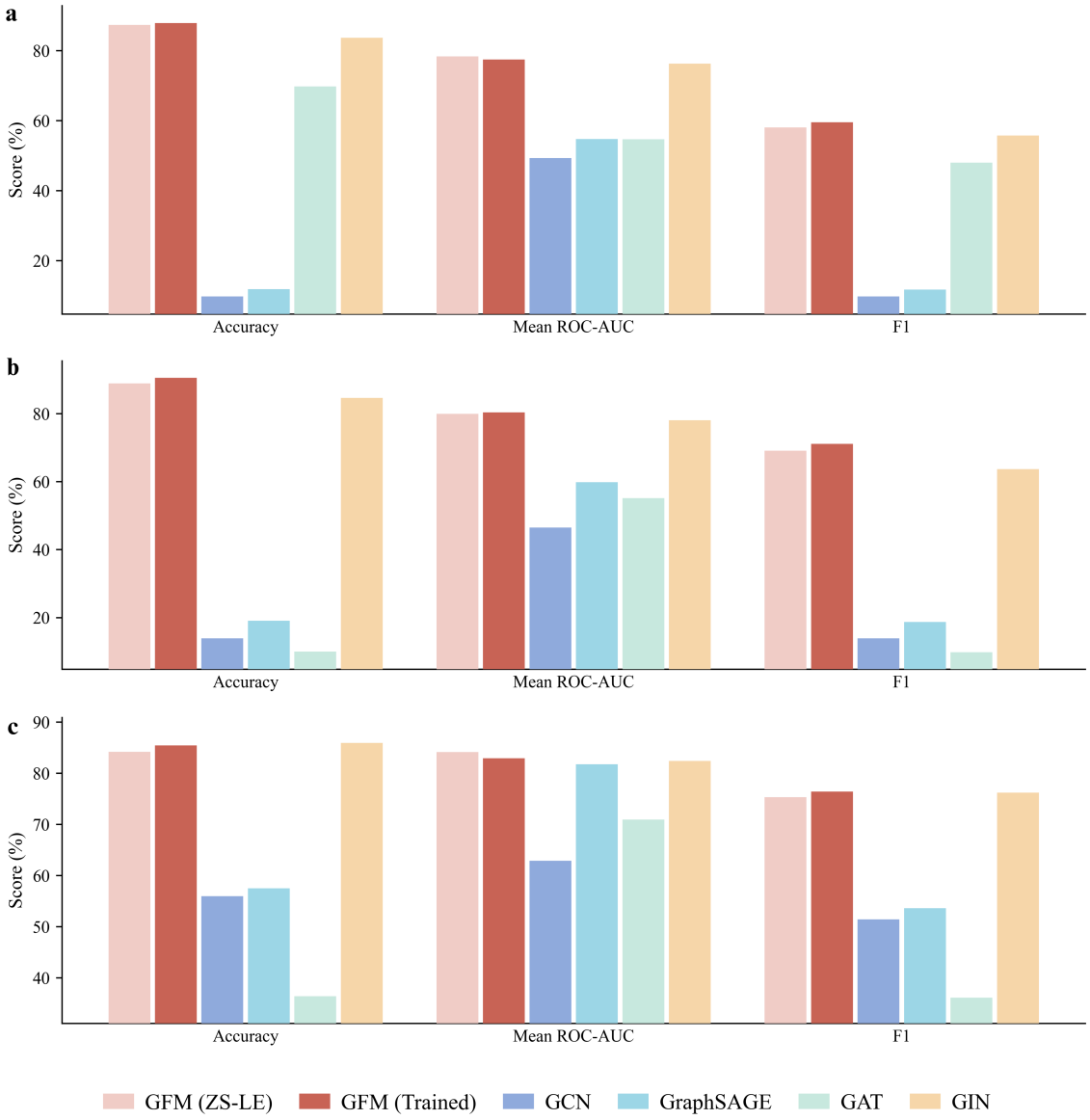

Fig. S16. GFM performance on the *C. elegans* STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the worm (taxon 6239) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed all supervised baselines on every metric across the three ontologies. Cross-species transfer from the predominantly mammalian and citation-graph pretraining distribution to an invertebrate proteome is preserved, with mean ROC-AUC gains over GIN of 7% to 11% across the three ontologies.

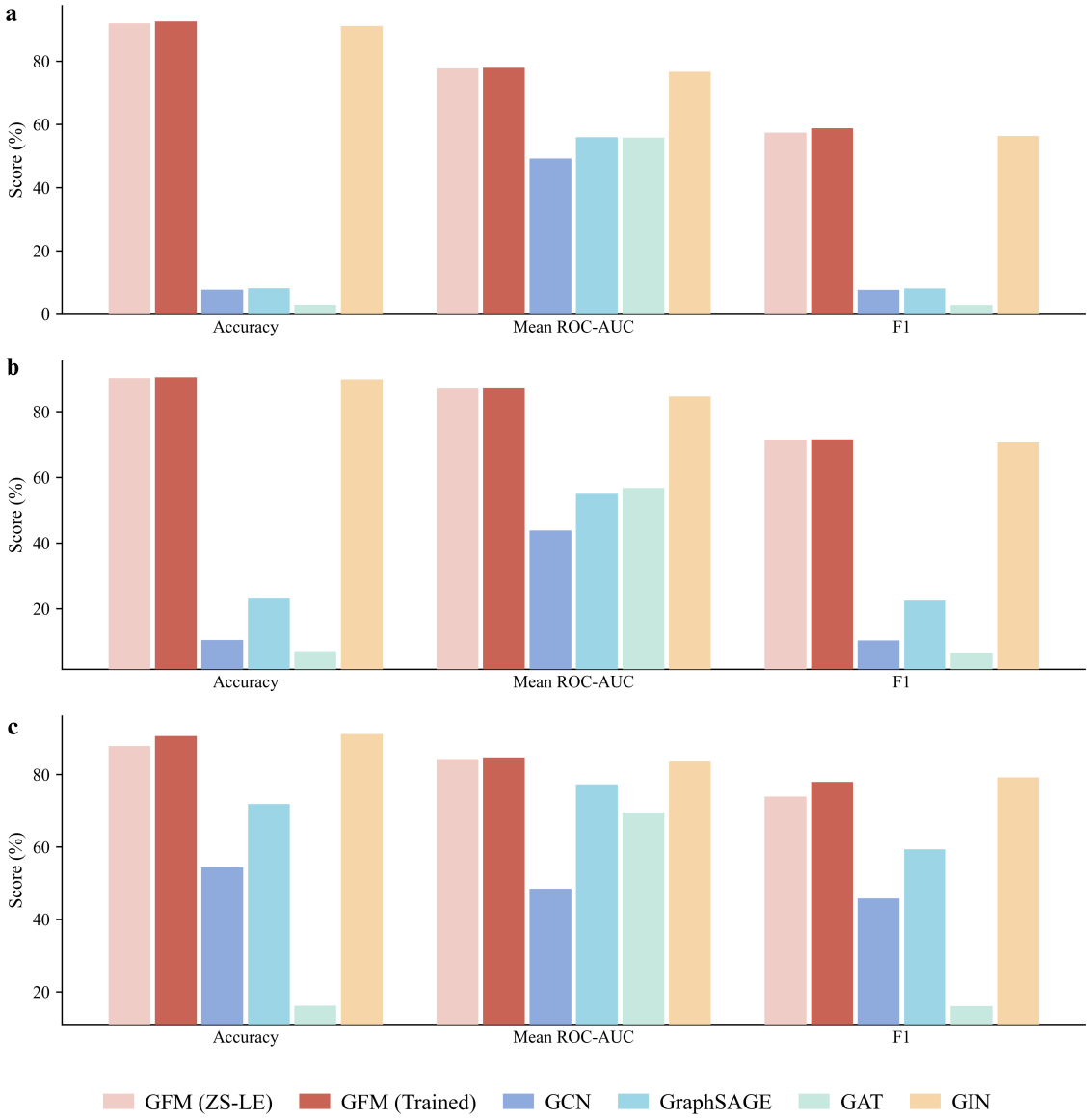

Fig. S17. GFM performance on the *Drosophila melanogaster* STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the fruit fly (taxon 7227) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed all four supervised baselines on accuracy and macro F1 across the three ontologies. GIN matches GFM (ZS-LE) on MF accuracy but is exceeded on every other metric in every panel. The result extends the cross-species generalization claim from vertebrates to a second invertebrate clade.

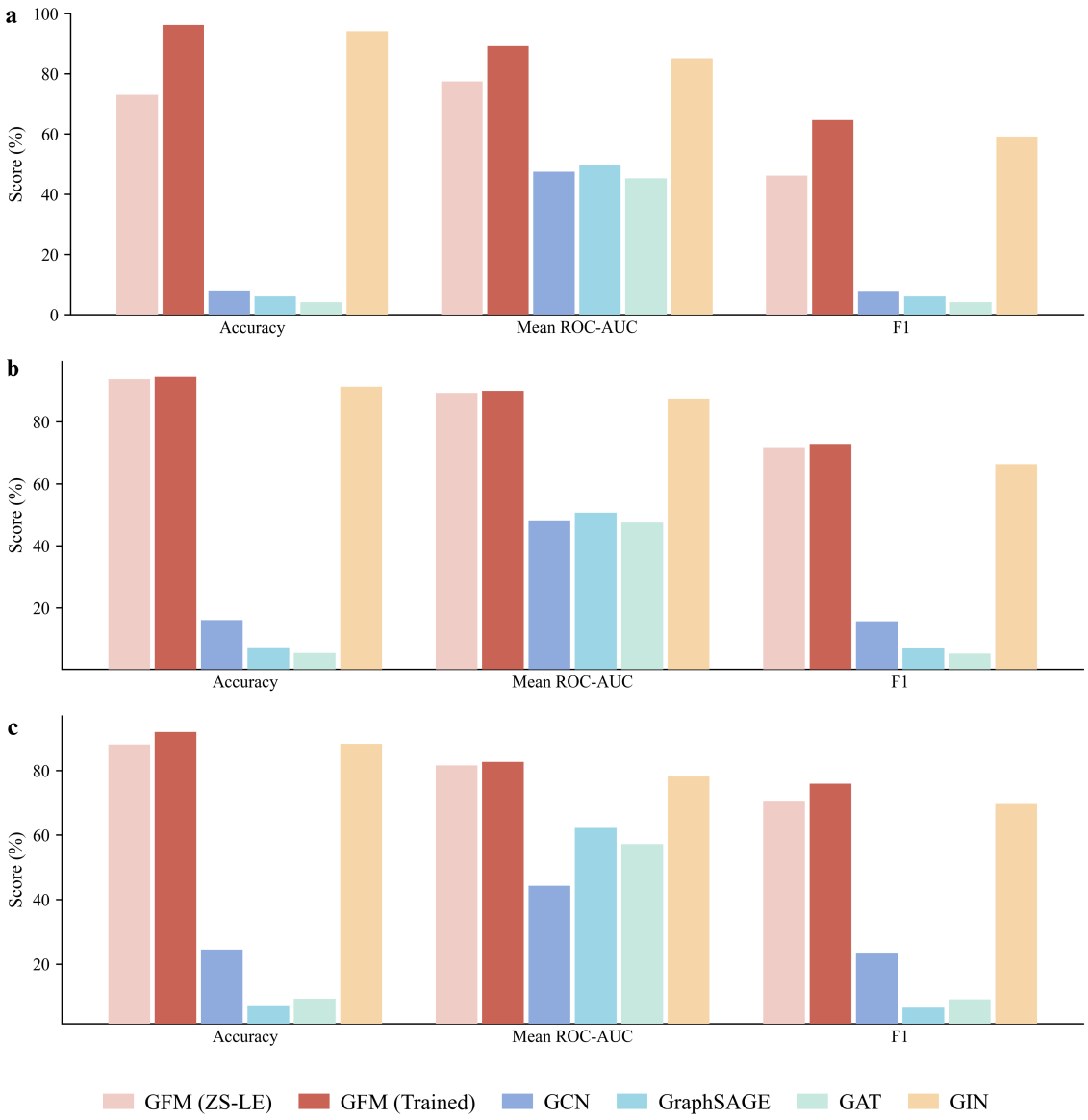

Fig. S18. GFM performance on the *Saccharomyces cerevisiae* STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the budding yeast (taxon 4932) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) match or exceed every supervised baseline on accuracy and macro F1 across the three ontologies. GIN is competitive on this network because the yeast proteome has relatively few GO terms passing the minimum-count filter (20 MF terms, 89 BP terms, 37 CC terms), reducing the effective task difficulty. The GFM variants remain the top performers in every panel.

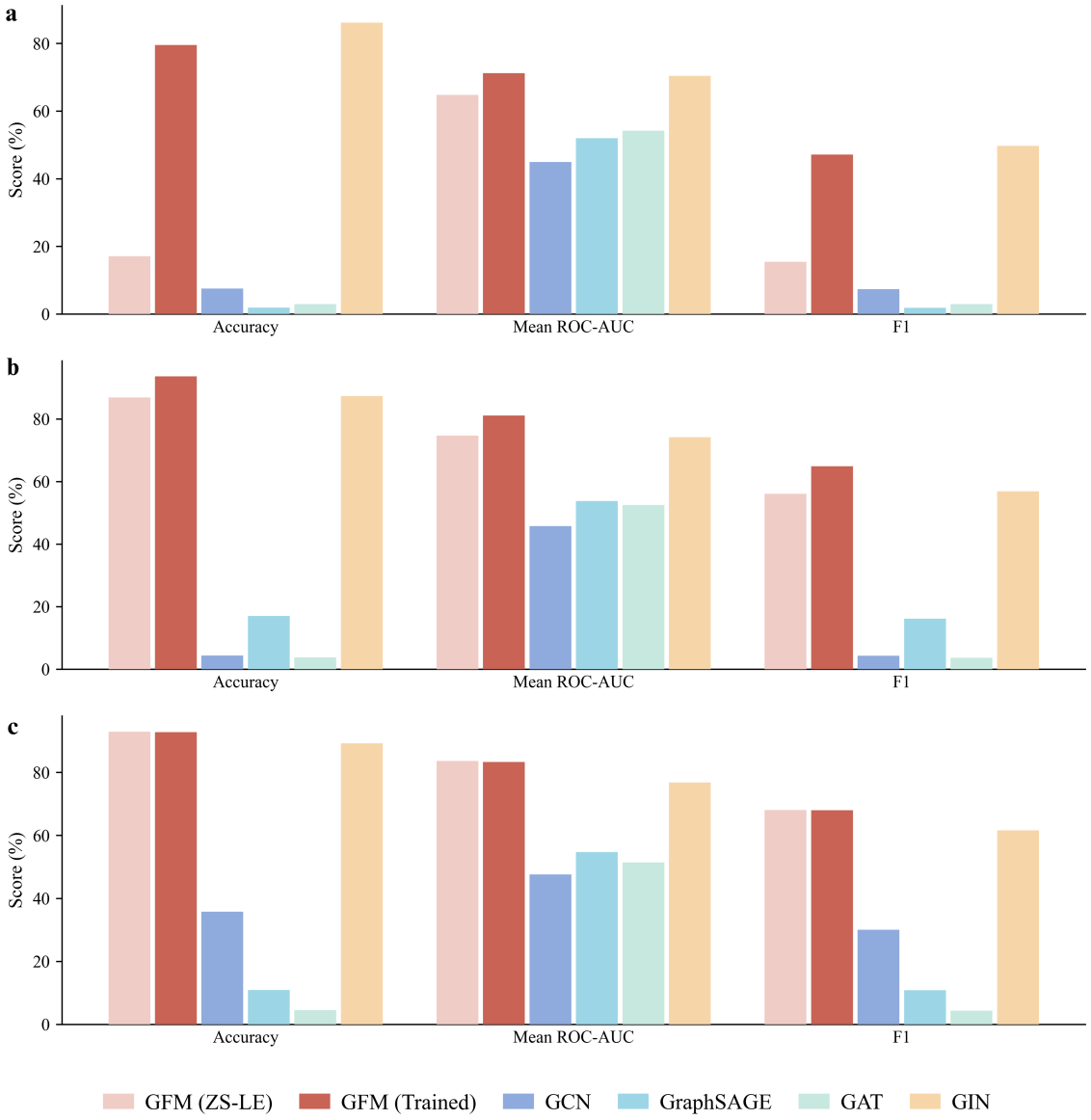

Fig. S19. GFM performance on the *Rattus norvegicus* STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the rat (taxon 10116) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (Trained) achieves the highest mean ROC-AUC on every ontology. GFM (ZS-LE) accuracy on BP (panel **b**) is lower than GIN because the zero-shot lightweight classifier head failed to converge on this particular task under the default training configuration, an effect specific to this single panel that does not affect the AUROC ranking. The macro F1 and AUROC patterns are consistent with the other mammalian STRING species.

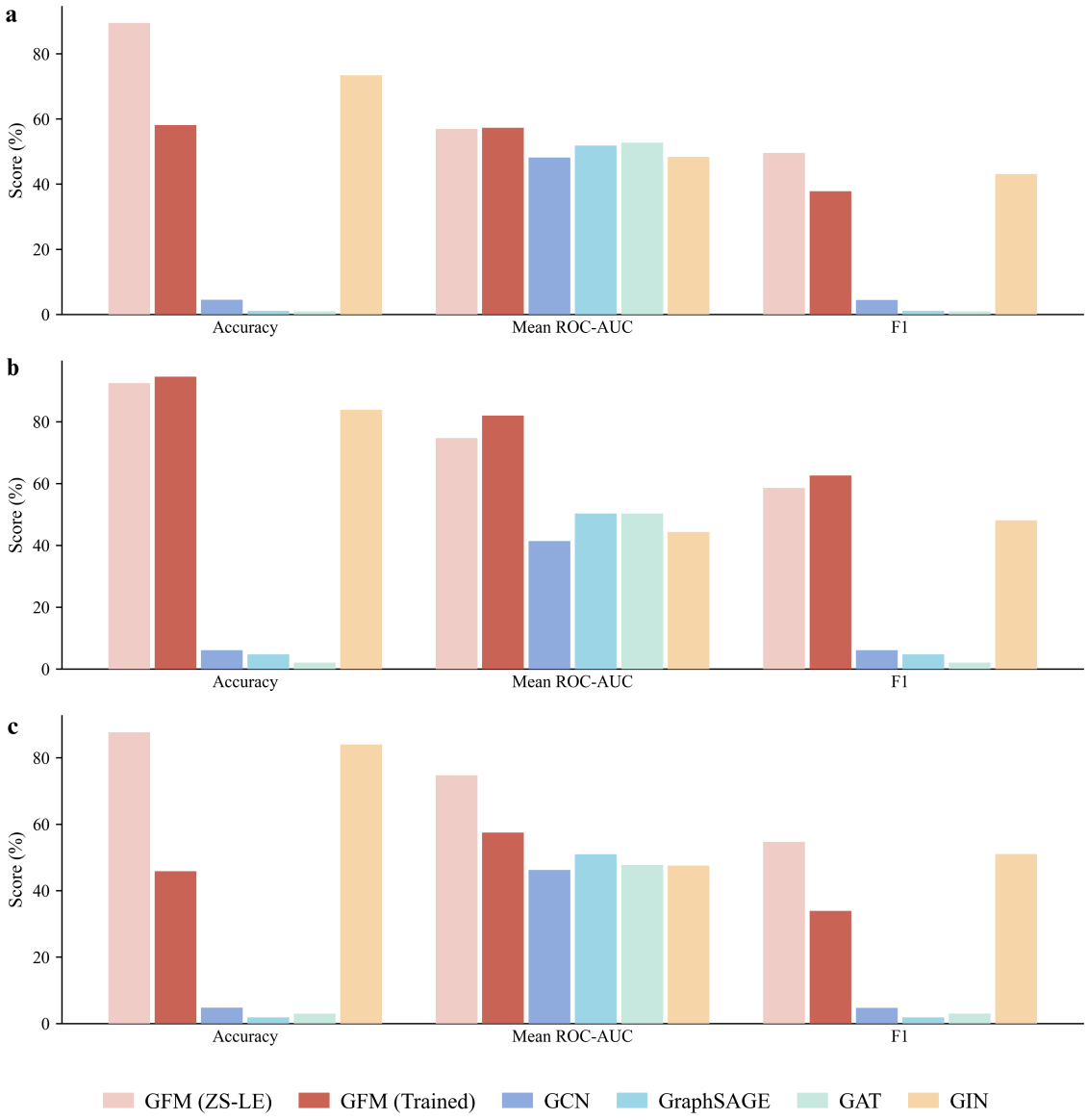

Fig. S20. GFM performance on the BioGRID human protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the BioGRID human interactome, an alternative human PPI source independent of STRING. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed all supervised baselines on every metric. The four message-passing baselines collapse to near-random AUROC (close to 50%) on all three ontologies, a pattern consistent with over-smoothing on dense human PPI graphs documented previously on StringGO. GFM (Trained) achieves 86.3%, 90.4%, and 82.0% mean ROC-AUC on MF, BP, and CC respectively, exceeding the strongest baseline by more than 21% on all three ontologies.

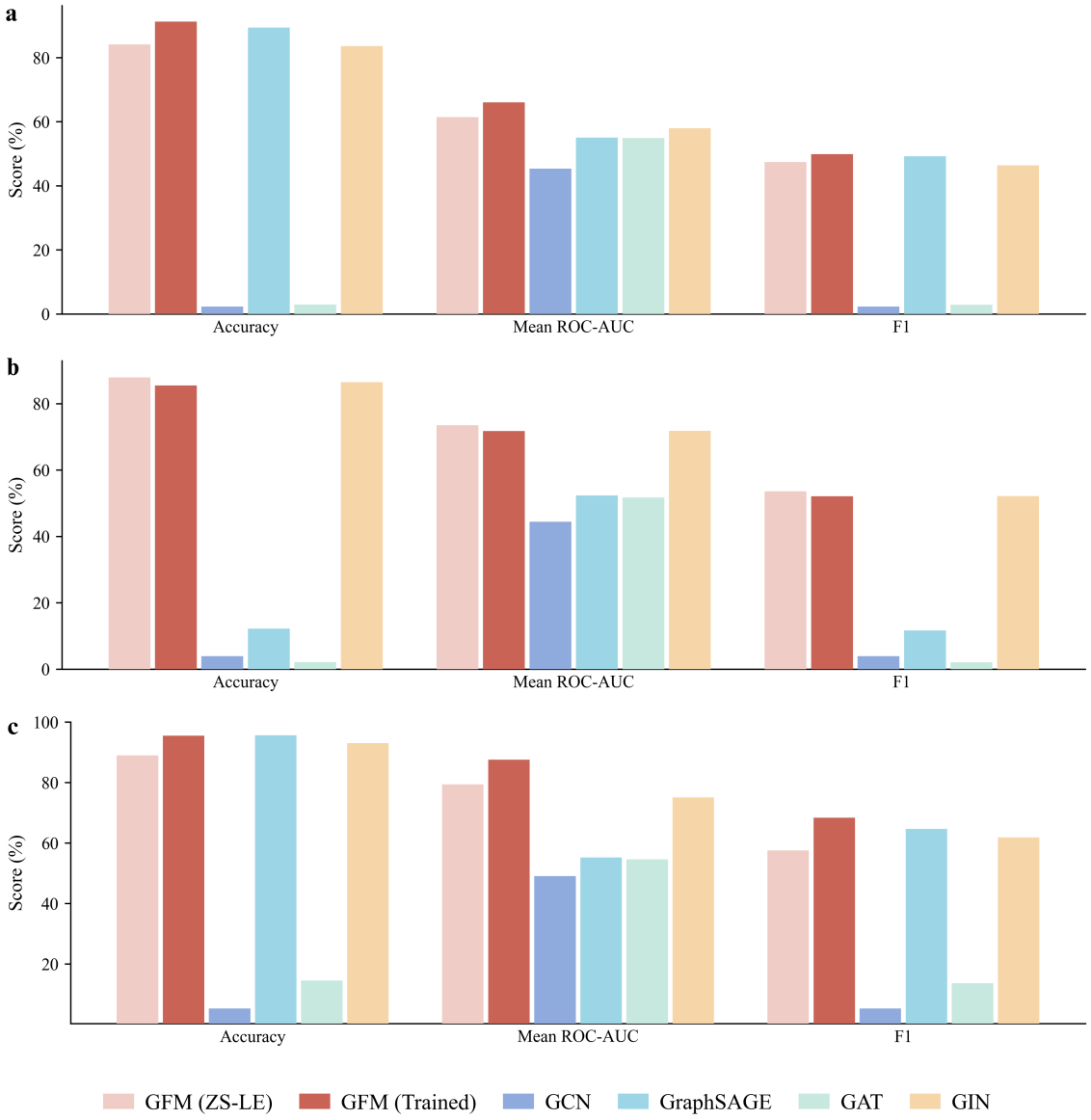

Fig. S21. GFM performance on the *Mus musculus* STRING protein interaction network across three GO ontologies. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on the mouse (taxon 10090) STRING v12 interactome. **a**, Molecular Function (MF) prediction. **b**, Biological Process (BP) prediction. **c**, Cellular Component (CC) prediction. GFM (ZS-LE) and GFM (Trained) exceed all supervised baselines on mean ROC-AUC across the three ontologies. GFM (ZS-LE) exceeds GFM (Trained) on AUROC for BP and CC, indicating that the frozen representation is already at or near the ceiling for these tasks on this species. GIN remains competitive on accuracy but trails on macro F1 and AUROC.

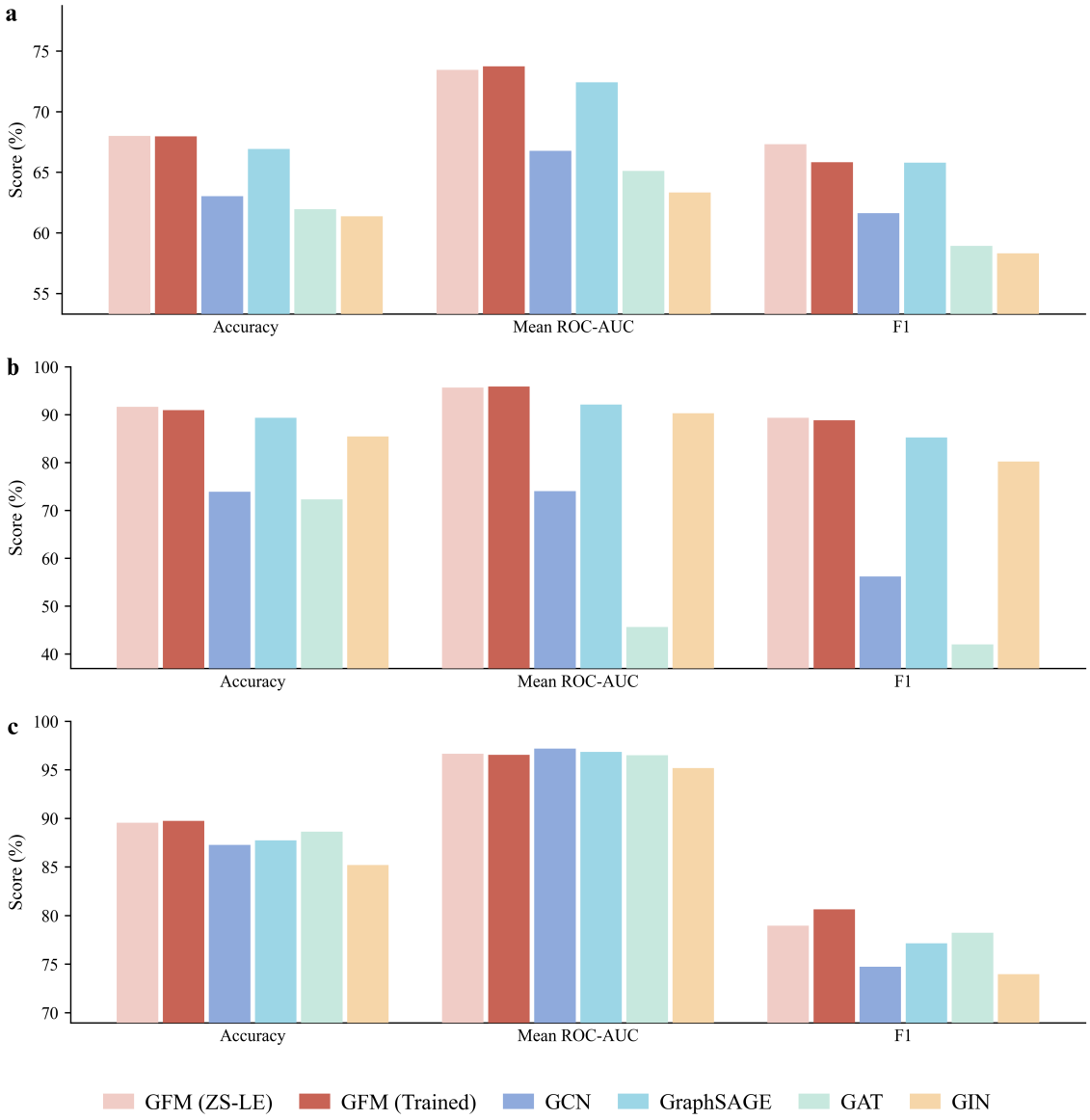

Fig. S22. GFM performance on three non-protein graph classification tasks. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on three single-label node classification benchmarks outside the protein interaction domain. **a**, DeezerEurope (musical preference prediction). **b**, TRRUST human transcription factor versus target binary classification. **c**, LastFMAsia (musical preference prediction). GFM (ZS-LE) and GFM (Trained) match or exceed the strongest supervised baseline on every panel. GFM (ZS-LE) exceeds all baselines on accuracy in panels **a** and **b**. The TRRUST binary task (panel **b**) tests whether the GFM encodes regulatory direction in addition to GO annotation structure. Performance across this heterogeneous collection of non-protein graphs indicates that the pretrained representations are not specific to PPI topology.

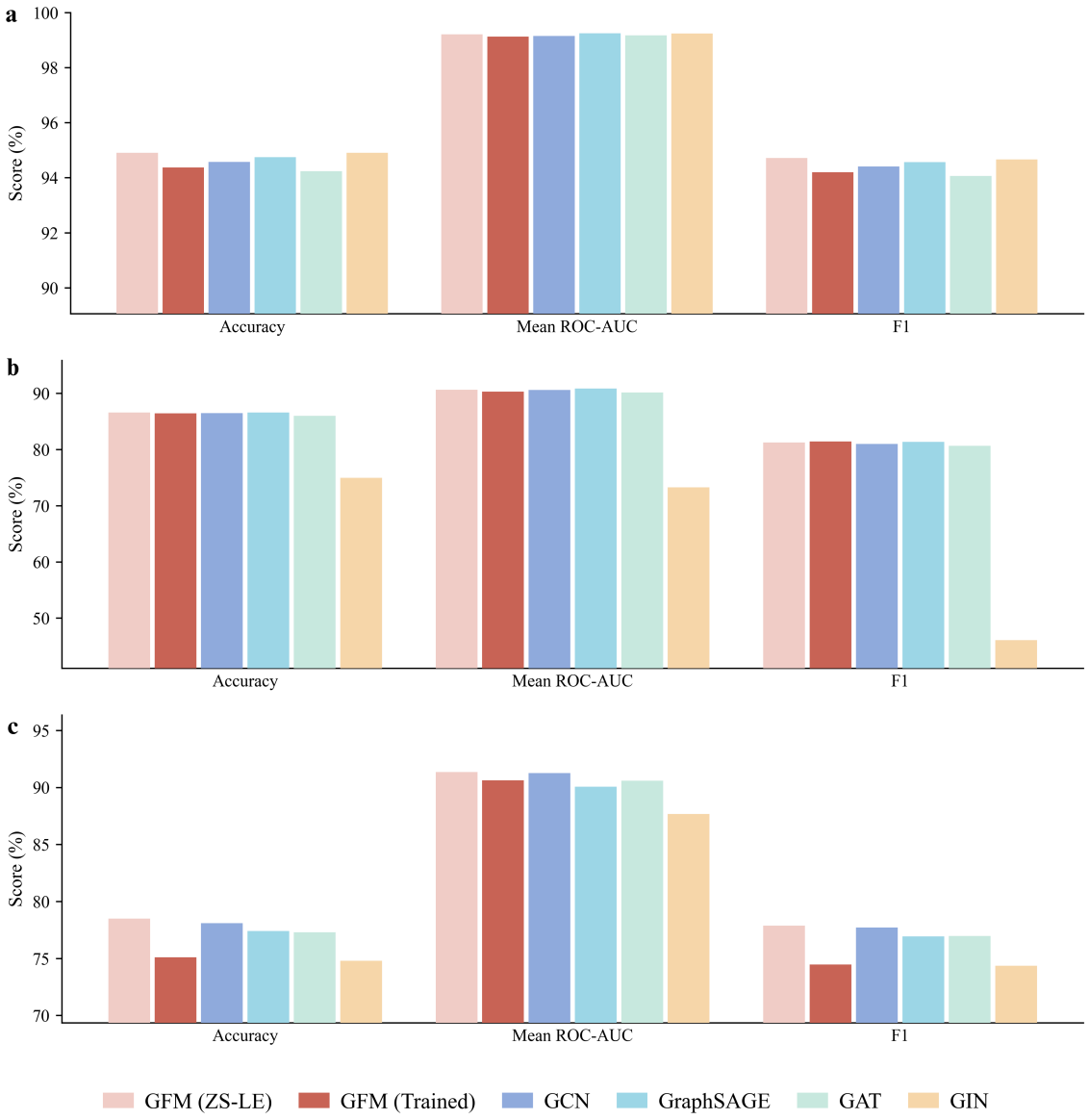

Fig. S23. GFM performance on three social and citation network classification tasks. Mean accuracy, mean ROC-AUC, and macro F1 are shown for all six models on three single-label node classification benchmarks drawn from social network and citation graph domains. **a**, Facebook Page-Page network (page category prediction). **b**, GitHub developer network (web or machine learning developer classification). **c**, PubMed citation network (paper topic classification). GFM (ZS-LE) and GFM (Trained) match the strongest supervised baseline within 2% on every panel, with neither GFM variant nor any baseline showing a decisive advantage. These three datasets fall within the citation, co-authorship, and social network families that constitute the pretraining distribution, so they test whether the GFM remains competitive on in-distribution graph families rather than testing transfer. The results confirm that GFM performance on the pretraining distribution remains comparable to specialized supervised baselines.