EVALUATING PROGRESS IN GRAPH FOUNDATION MODELS: A COMPREHENSIVE BENCHMARK AND NEW INSIGHTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph foundation models (GFM) aim to acquire transferable knowledge by pre-training on diverse graphs, which can be adapted to various downstream tasks. However, domain shift in graphs is inherently two-dimensional: graphs differ not only in what they describe (*topic* domains) but also in how they are represented (*format* domains). Most existing GFM benchmarks vary only topic domains, thereby obscuring how knowledge transfers across both dimensions. We present a new benchmark that jointly evaluates topic and format gaps across the full GFM pipeline, including multi-domain pre-training and few-shot downstream adaptation, and provides a timely evaluation of recent GFMs in the rapidly evolving landscape. Our protocol enables controlled assessment in four settings, covering seen and unseen datasets as well as cross-topic and cross-format evaluation, designed to disentangle topic generalization from robustness to format shifts. We conduct extensive evaluations of eight state-of-the-art GFMs on 33 datasets spanning seven topic domains and six format domains, surfacing new empirical results and insights for future research. Code and data are available at `https://anonymous.4open.science/r/GFMBenchmark-7EA0`.

1 INTRODUCTION

Foundation models have emerged as a unifying paradigm in modern artificial intelligence, where models are pre-trained on broad and heterogeneous data and subsequently adapted to diverse downstream tasks (Achiam et al., 2023; Touvron et al., 2023). Their success demonstrates the potential of large-scale pre-training to learn general-purpose representations that transfer across tasks and data distributions. Recent studies have extended this paradigm to graph-structured data, giving rise to *graph foundation models* (GFM) (Liu et al., 2025a; Mao et al., 2024). By pre-training on graph datasets from multiple sources or domains, GFMs aim to learn transferable structural and semantic knowledge that generalizes to unseen tasks, graphs, or domains with limited task-specific labels.

However, benchmarking GFMs can be complex and challenging, since domain shift in graphs is inherently multi-faceted. Graph heterogeneity manifests not only in what graphs represent but also in how they are structured and represented. Inspired by Wettig et al. (2025), we categorize graph domains into *topic* and *format* domains. From the semantic perspective, graphs arise from diverse **topic domains**, including citation networks (Radicchi et al., 2011; McLaren & Bruner, 2022), social networks and the Web (Freeman et al., 2004; Marin & Wellman, 2011), e-commerce graphs (Hu et al., 2020b; Zhao et al., 2020), molecular graphs (Kearnes et al., 2016; You et al., 2018), protein interaction networks (Giuliani et al., 2008; Borgwardt et al., 2005), and financial networks (Bardoscia et al., 2021; Haldane, 2013). These domains differ in node/edge semantics and feature spaces, creating pronounced semantic gaps that challenge multi-domain pre-training and cross-domain transfer. Beyond semantics, graphs also vary widely in **format domains**, including homogeneous vs. heterogeneous graphs (Zhang et al., 2019; Wang et al., 2022), homophilic vs. heterophilic structures (Rossi et al., 2024; Bi et al., 2024), static vs. dynamic graphs (Harary & Gupta, 1997; Manessi et al., 2020), knowledge graphs (Chen et al., 2020), and text-attributed graphs (Zhao et al., 2023; Wen & Fang, 2023). These format differences impose distinct modeling assumptions across topological patterns, temporal dependencies, and attribute or relational schemas, which can fundamentally affect both pre-training and adaptation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite the variety across topics and formats, existing GFM benchmarks typically adopt a one-dimensional notion of domain diversity (Yang et al., 2025; Xu et al., 2024; Liu et al., 2025b; Chen et al., 2024b). Specifically, they vary topic domains while keeping the graph format fixed (Yang et al., 2025; Xu et al., 2024; Chen et al., 2024b), or restrict evaluation to a narrow task or application area (Liu et al., 2025b; Platonov et al., 2026). More recent efforts focus on curvature-stratified evaluation (Wang et al., 2026) and Prior-data Fitted Network (PFN)-based GFMs (Platonov et al., 2026), complementing our focus on jointly assessing topic and format shifts. Without explicitly disentangling topic and format domains, it remains unclear whether observed transfer performance reflects generalization across both dimensions, potentially conflating topics and formats and limiting a comprehensive understanding of current GFMs. Beyond dataset design, existing benchmarks are also limited in method coverage: they often evaluate a specific type of methods (Liu et al., 2025b) or tend to lag behind the rapidly evolving literature (Yang et al., 2025; Xu et al., 2024; Chen et al., 2024b). As a result, conclusions from prior benchmark studies may not extrapolate to a broader range of methods or more recent GFMs.

To address these gaps, we present a new GFM benchmark that explicitly accounts for both topic and format domains in pre-training and downstream adaptation, while covering representative GFMs across major modeling regimes. Specifically, we construct unified evaluation settings by performing multi-domain pre-training across diverse *topics* and *formats* and adapt the resulting models to downstream tasks on both *seen* and *unseen* graphs. By controlling the topic and format facets, our benchmark enables fine-grained analysis of cross-topic and cross-format transfer. This two-axis setup, together with broader method coverage, enables different forms of generalization and their interactions to be systematically characterized.

In summary, we make the following contributions. (1) We formalize a two-dimensional view of graph domains, distinguishing their *topics* from *formats*. (2) We construct a comprehensive benchmark that spans diverse topic and format domains and includes a broad range of representative GFMs. (3) We introduce unified evaluation settings that separately and jointly assess the effect of topic and format shifts in both seen and unseen downstream targets. (4) We provide extensive empirical analyses that reveal previously uncharacterized generalization behaviors and limitations of existing GFMs, offering actionable insights for future GFM design.

2 Related Work

Graph foundation models. Existing GFMs can be broadly grouped by how they mitigate domain gaps in graphs. Several GFMs (Yu et al., 2024b; Zhao et al., 2024; Wang et al., 2025) target feature mismatch by introducing domain-conditioned tokens and prompts, shared coordinators, or lightweight adapters that allow a single backbone to operate across heterogeneous input spaces. Another line of GFMs focuses on structural mismatch across domains. Existing approaches align topology statistics or modulate message passing with structure-aware mechanisms (Yu et al., 2025a; Wang et al., 2025; 2024; Zhao et al., 2024), while more recent methods explore Riemannian geometry (Sun et al., 2025; 2026), random walks (Kim et al., 2026), structure-aware semantic augmentation (Shi et al., 2026), and synthetic structures for prior-data fitted networks (Eremeev et al., 2026). Collectively, these methods aim to reduce domain-dependent structural biases and improve transfer across graphs with different structural regimes. Additionally, some GFMs (Zhu et al., 2025; Wen & Fang, 2023; He et al., 2025; Wang et al., 2024) leverage textual and multimodal signals as an external semantic anchor, aligning graph representations with text or other modalities via contrastive or prompt-based objectives. While effective when such signals are available, their generalization to universal graph datasets is limited. Another orthogonal direction uses explicit supervision in pre-training (Liu et al., 2024; Huang et al., 2023; Tang et al., 2024a;b; Xia et al., 2024; Yuan et al., 2025; Chen et al., 2024a; Li et al., 2024; Zhao et al., 2025) with labeled source-domain data, which diverges from the traditional pre-training paradigm, and thus falling outside the scope of this benchmark. We summarize representative GFMs in Table 7, with further details in Appendix A.3.

Benchmarks for graph foundation models. Existing GFM benchmarks characterize domain diversity along a single axis. They typically vary topic domains while implicitly fixing (or tightly constraining) the graph format (Yang et al., 2025; Xu et al., 2024; Liu et al., 2025b; Chen et al., 2024b). This one-dimensional design makes results hard to attribute: observed gains may stem from semantic or structural generalization, robustness to format changes, or simply a favorable match

between the pre-training and downstream formats, but these factors are not separated. Beyond this shared limitation, existing benchmarks also differ in scope and protocol. Yang et al. (2025) primarily focuses on heterogeneous graphs and performs pre-training on a single dataset, which differs from the standard GFM setting of multi-domain pre-training. Xu et al. (2024) primarily benchmarks graph self-supervised learning on a small set of citation and social datasets, rather than the standard GFM pipeline of multi-domain pre-training followed by cross-domain adaptation. Liu et al. (2025b) provides a careful study of robust fine-tuning, but its focus on molecular graphs limits the conclusions to a single topic domain. Chen et al. (2024b) targets text-attributed graphs where LLM-derived embeddings help unify feature spaces, and is not designed to evaluate broader graph formats. More recently, CurvBench (Wang et al., 2026) examines curvature-stratified evaluation, comparing GFMs across curvature regimes to expose geometry-dependent performance differences, whereas Platonov et al. (2026) evaluate PFN-based methods such as GraphPFN (Eremeev et al., 2026), G2T-FM (Eremeev et al., 2025), and TAG (Hayler et al., 2025) for the node property prediction task. Our benchmark complements these efforts by treating domain shift as inherently two-dimensional, across both semantic topics and graph formats, and evaluating GFMs under a two-stage pipeline of multi-domain pre-training followed by cross-domain few-shot adaptation.

3 Target Models, Domain Composition and Datasets

In this section, we introduce the evaluated GFMs, define a two-dimensional notion of graph domains (topic vs. format), and describe the benchmark datasets.

Evaluated GFMs. We consider GFMs that operate in two stages: (1) self-supervised pre-training on a broad multi-domain corpus of graphs, and (2) cross-domain few-shot adaptation to diverse downstream graphs. An overview of the two-stage pipeline is provided in Appendix A.1. Descriptions of the evaluated GFMs following this pipeline (Yu et al., 2024b; 2025a; Wang et al., 2025; Zhao et al., 2024; Zhu et al., 2025; He et al., 2025; Wen & Fang, 2023; Wang et al., 2024) are provided in Appendix A.3. Other GFM-style approaches, particularly those requiring labeled source-domain data during pre-training, fall *outside* the scope of this benchmark.

Domain composition. A central design goal of our benchmark is to make graph domain explicit along two largely orthogonal axes: topics and formats. Topic domain describes what a graph represents (semantics), whereas format domain describes how the graph is represented (schema). Table 1 summarizes the domains and datasets used in our benchmark, with details given in Appendix A.4.

Topic domains. We define a topic domain by the semantic meaning of nodes and edges and the typical downstream applications. Thus, shifts in topic domains manifest as changes in node and edge semantics (e.g., papers and citations versus users and interactions), feature distributions, and downstream label semantics and distributions. Our benchmark covers a broad range of common topics in graph datasets, including citation networks, social networks and the Web, e-commerce graphs, financial networks, common-sense graphs, molecular graphs, and protein interaction networks.

Format domains. The format domain captures the schema or representational form of a graph and therefore induces different modeling assumptions over topological patterns, temporal dependencies and attribute or relational schemas. Concretely, we consider variations such as homogeneous vs. heterogeneous graphs, homophilic vs. heterophilic graphs, static vs. dynamic graphs, relational graphs, and text-attributed graphs. These format differences influence core components of graph learning, such as message passing behavior, temporal dynamics modeling, and relational reasoning.

Datasets. Topic and format are conceptually orthogonal: Graphs may share topic semantics but adopt different formats. For instance, *Cora* (Liu et al., 2024) and *DBLP* (Wang et al., 2022) both fall under citation networks, yet *Cora* is typically represented as a homogeneous graph while *DBLP* is heterogeneous with typed nodes and edges. Conversely, graphs may share the same format while originating from different topics: *Actor* (Yu et al., 2025c) and *HIV* (Liu et al., 2024) are both homogeneous graphs, but belong to social and molecule topic domains, respectively.

4 Benchmarking Protocols

This section outlines our benchmark setup and four evaluation settings for probing GFM generalization across topic and format domains.

Table 1: Summary of the datasets by topics and formats.

Topic	Dataset	# Graphs	Avg. # nodes	Avg. # edges	Format					Used for	Task		
					Hetero.	Homophily	Rel.	Dyn.	Text	pre-train	NC	EC	GC
Citation	Cora (Liu et al., 2024)	1	2,708	10,556	×	0.81	×	×	✓	✓	✓	×	×
	Pubmed (Liu et al., 2024)	1	19,717	44,338	×	0.80	×	×	✓	×	✓	×	×
	ogbn-arxiv (Hu et al., 2020a)	1	169,343	1,166,243	×	0.66	×	×	✓	✓	×	×	×
	ogbn-mag (Hu et al., 2020a)	1	1,939,743	21,111,007	✓	0.30	×	×	×	×	✓	×	×
	ACM (Tang et al., 2024b)	1	10,942	547,872	✓	0.88	×	×	×	✓	✓	×	×
	DBLP (Wang et al., 2019)	1	26,128	239,566	✓	-	×	×	×	✓	×	×	×
Social Web	Reddit (Yu et al., 2025b)	1	10,984	672,447	×	-	×	✓	×	✓	✓	×	×
	Wikipedia (Yu et al., 2025b)	1	9,227	157,474	×	-	×	✓	×	×	✓	×	×
	Actor (Li et al., 2025)	1	4,416	12,172	×	0.56	×	×	✓	×	✓	×	×
	Texas (Li et al., 2025)	1	187	328	×	0.12	×	×	✓	✓	×	×	×
	Wisconsin (Li et al., 2025)	1	265	530	×	0.20	×	×	✓	✓	✓	×	×
	Chameleon (Rozemberczki et al., 2021)	1	2,277	36,101	×	0.24	×	×	×	×	✓	×	×
	Cornell (Li et al., 2025)	1	195	304	×	0.13	×	×	✓	✓	✓	×	×
	IMDB (Wang et al., 2019)	1	21,420	86,642	✓	-	×	×	×	✓	×	×	×
E-commerce	Photo (Yu et al., 2024b; 2025a)	1	7,650	238,162	×	0.83	×	×	×	✓	✓	×	×
	Computers (Yu et al., 2024b; 2025a)	1	13,752	491,722	×	0.78	×	×	×	✓	×	×	×
	Amazon (Cen et al., 2019)	1	10,099	148,659	✓	-	×	×	×	✓	×	×	×
	Amazon-HeTGB (Li et al., 2025)	1	24,492	93,050	×	0.38	×	×	✓	✓	×	×	×
	Products (Chen et al., 2024a)	1	2,449,029	61,859,140	×	0.81	×	×	✓	×	✓	×	×
Common sense	Wiki (Huang et al., 2023)	1	4,815,483	20,624,575	×	-	✓	×	×	×	×	✓	×
	FB15K-237 (Huang et al., 2023)	1	14,505	310,079	×	-	✓	×	×	✓	×	✓	×
	NELL (Huang et al., 2023)	1	65,755	234,355	×	0.83	✓	×	×	✓	×	×	×
	WN18RR (Liu et al., 2024)	1	40,559	92,583	×	-	✓	×	×	×	×	✓	×
Finance	T-Finanace (Tang et al., 2022)	1	39,357	21,222,543	×	0.97	×	×	×	×	✓	×	×
	Elliptic (Weber et al., 2019)	1	203,769	234,355	×	0.71	×	×	×	✓	✓	×	×
	DGraph (Huang et al., 2022)	1	3,700,550	4,300,999	×	0.37	×	×	×	×	✓	✓	×
Molecule	HIV (Liu et al., 2024)	41,120	25.5	27.5	×	-	×	×	✓	✓	×	×	✓
	PCBA (Liu et al., 2024)	437,927	26.0	28.1	×	-	×	×	✓	×	×	×	✓
	COX2 (Morris et al., 2020)	467	41.2	43.4	×	-	×	×	×	✓	×	×	✓
	BZR (Morris et al., 2020)	405	35.8	38.4	×	-	×	×	×	×	×	×	✓
Protein	PROTEINS (Morris et al., 2020)	1,113	39.06	72.82	×	-	×	×	×	✓	×	×	✓
	ENZYMES (Morris et al., 2020)	600	32.63	62.14	×	-	×	×	×	✓	×	×	×
	ogbn-proteins (Hu et al., 2020a)	1	132,534	79,122,504	×	-	×	×	×	×	✓	×	×

The *Used for pre-train* column indicates whether the dataset is used for pre-training. The *Task* column specifies the supported downstream tasks (NC: node classification, EC: edge classification, and GC: graph classification).

Benchmark setup. (1) *Data pre-processing.* To enable fair and reproducible comparisons, we standardize all datasets into a unified graph interface while preserving task- and format-specific semantics whenever supported by a method. This includes consistent treatment of graph direction, node and edge types, temporal information, relation attributes, textual fields, and feature dimensions. The complete data object specification, format-specific transformations, feature alignment, and dataset-specific adaptations are provided in Appendix B.1. (2) *Downstream tasks and data splits.* We evaluate GFMs on node, edge, and graph classification using unlabeled pre-training datasets and one- and five-shot downstream adaptation. Following the standard K-shot protocol (Liu et al., 2023; Yu et al., 2024a), we report the mean and standard deviation of Accuracy and Macro-F1 over repeated task construction and random seeds. The complete data-split, episode-sampling, and task-specific batching procedures are provided in Appendix B.2. (3) *Implementation details.* We use the authors' official open-source implementations whenever available. Hyperparameters are tuned within the search spaces recommended in the original papers. Full implementation details and hyperparameter configurations are deferred to Appendix B.3.

Evaluation settings. We design four evaluation settings to probe GFM generalization under seen and unseen data, as well as topic and format shifts (Fig. 4, Appendix A.2).

Setting I: Pre-train on diverse topics and formats; adapt to unseen datasets. We construct $\mathcal{D}_{\text{pre}}$ by selecting unlabeled datasets spanning a broad range of topic and format domains, and hold out a disjoint set $\mathcal{D}_{\text{unseen}}$ such that no graph in $\mathcal{D}_{\text{down}}$ is used during pre-training, i.e., $\mathcal{D}_{\text{down}} = \mathcal{D}_{\text{unseen}}$ and $\mathcal{D}_{\text{pre}} \cap \mathcal{D}_{\text{unseen}} = \emptyset$. Concretely, we use all those indicated under the *Used for pre-train* column in Table 1 as $\mathcal{D}_{\text{pre}}$, and treat the remaining datasets as $\mathcal{D}_{\text{unseen}}$.

Setting II: Pre-train on diverse topics and formats; adapt to seen datasets. Downstream evaluation is performed on the same datasets used for pre-training, i.e., $\mathcal{D}_{\text{down}} = \mathcal{D}_{\text{pre}}$, while task labels $\mathcal{Y}_{\text{down}}$ are available only at the downstream stage. We adopt the same $\mathcal{D}_{\text{pre}}$ as in Setting I.

Setting III: Pre-train on single-topic domains; adapt across topics. To isolate the effects of topic domains, we pre-train separate models on three topics: *citation*, *social*, and *molecule*. Each model is

Table 2: Accuracy of one-shot node classification on unseen datasets.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCN	48.62±8.40	5.83±0.70	48.02±39.11	26.50±11.82	26.71±4.33	21.12±5.67	42.16±39.45	32.88±15.35	51.54±13.11
GAT	45.50±9.29	4.86±0.60	45.32±24.72	23.90±11.13	26.24±4.27	22.11±6.28	53.86±38.15	33.46±16.13	47.93±14.10
SIMPLE-HGN	/	**9.08**±1.10	/	/	/	/	/	/	/
FAGCN	/	/	/	20.02±6.13	23.79±3.44	/	/	36.69±18.43	/
GCOPE	41.85±8.34	0.27±0.09	43.23±25.98	24.16±6.83	21.66±3.41	OOT	45.42±22.22	OOT	45.01±12.73
MDGPT	40.50±7.33	5.94±0.63	51.02±25.40	**31.72**±8.93	28.36±5.04	15.99±3.48	58.95±23.03	37.52±12.02	50.72±10.78
MDGFM	39.50±7.20	2.93±1.05	44.25±28.21	19.80±7.00	25.95±3.75	8.11±2.17	57.07±19.12	36.08±24.19	50.28±8.80
SAMGPT	49.23±8.07	5.32±0.68	**56.50**±27.26	27.09±7.73	**28.65**±5.12	**22.68**±4.03	58.28±15.86	39.42±13.36	**52.25**±11.97
G2P2	48.56±6.77	-	38.62±10.45	29.91±6.81	-	-	**63.04**±18.64	38.30±10.71	-
GRAPHCLIP	36.67±4.26	-	34.67±21.87	25.18±5.87	-	-	44.77±21.85	**39.63**±26.42	-
GFT	**53.56**±10.27	1.50±0.29	55.00±17.19	25.69±5.93	27.32±4.08	11.36±4.49	48.43±23.94	35.24±12.68	51.03±7.93
UNIGRAPH2	45.27±9.41	2.21±1.19	52.23±30.37	27.35±9.23	26.02±5.71	22.61±4.70	58.59±22.94	36.37±13.57	50.06±6.42

OOT: out-of-time, execution exceeding three days. OOM: GPU out-of-memory. G2P2 and GRAPHCLIP require textual labels for downstream tasks; marked as '-' when unavailable. Format-specific methods, SIMPLE-HGN and FAGCN, are evaluated only on graphs of their intended formats; marked with '/' otherwise.

Best/runner-up results are in bold/underlined. **The same notations apply to subsequent tables.**

adapted to the same downstream datasets as in Setting I, enabling controlled comparisons between same- and cross-topic adaptations, as well as against broad multi-topic pre-training.

Setting IV: Pre-train on a base format; adapt to other formats. To isolate the effects of format domains, we pre-train on datasets in a canonical base format, corresponding to homogeneous, homophilic, static, single-relation, text-free graphs listed in Table 1, among those marked under *Used for pre-train*. We then adapt and evaluate the model on downstream datasets in various other formats.

5 EMPIRICAL RESULTS AND ANALYSIS

We present and analyze the evaluation results across the four evaluation settings, each addressing a distinct research question (RQ). For brevity, here we only present one-shot node classification performance, with detailed results provided in Appendices C.1–C.5, including more metrics, edge and graph classification performance, five-shot results, and efficiency where applicable.

5.1 EMPIRICAL RESULTS AND OBSERVATIONS

RQ1: *After multi-domain pre-training over diverse topic and format domains, can GFMs reliably adapt to unseen downstream datasets?* We report one-shot classification performance on unseen datasets for **Setting I** in Tables 2 and 10. Besides GFMs, we include conventional supervised GNNs (GCN and GAT) as baselines, which are trained directly on downstream data without pre-training. Note that conventional graph pre-training methods that pre-train and evaluate on the same dataset do not apply to the unseen downstream setting considered here.

We first observe that no single GFM dominates across unseen datasets. Depending on the dataset and task, SAMGPT, MDGPT, GFT and MDGFM are among the most competitive methods. When textual labels are available, G2P2 and GRAPHCLIP tend to perform well due to their explicit use of textual labels for classification. For instance, G2P2 excels on *T-Finance*, and GRAPHCLIP leads on *DGraph* in node classification. A notable exception is the heterogeneous *ogbn-mag* dataset, where the format-specific SIMPLE-HGN achieves the best results, outperforming both general supervised baselines and all evaluated GFMs. Second, while GFMs largely outperform conventional supervised GNNs, the improvements are inconsistent. On certain datasets (e.g., *DGraph* and *T-Finance*), GFM yields clear improvements over GCN and GAT. However, on some datasets, improvements are limited, with supervised GNNs trained from scratch remaining competitive (e.g., *ogbn-mag*).

RQ2: *After multi-domain pre-training over diverse topic and format domains, can GFMs reliably adapt to seen downstream datasets?* We report one-shot performance on seen datasets for **Setting II** in Tables 3 and 21. Similar to RQ1, we also compare with *supervised* GCN and GAT. Additionally, we compare with *conventional pre-training methods*, which are pre-trained on the downstream graph without any labels and then adapted to the same graph with task-specific labels. Since these methods require access to the downstream graph during pre-training, they apply only to the *seen* setting here, but not the *unseen* setting in RQ1. Such methods include general-purpose approaches (DGI and GRAPHPROMPT), and where applicable, format-specific approaches (HECO for heterogeneous graphs, TGN and DDGCL for dynamic graphs, and DSSL for heterophilic graphs).

Table 3: Evaluation of one-shot node classification on seen datasets.

Methods	Cora		ACM		Reddit		Wisconsin		Photo		Elliptic	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	41.76±10.38	37.85±9.98	38.31±8.08	24.12±10.36	50.90±33.56	32.75±16.73	34.20±7.94	22.49±4.55	53.87±12.35	52.05±11.90	41.74±18.55	32.74±11.50
GAT	43.48±9.64	39.92±8.96	35.82±5.75	22.37±8.94	48.68±30.94	32.30±15.74	34.18±8.57	23.06±5.94	46.38±11.11	45.49±9.65	38.73±20.26	29.95±11.87
SIMPLE-HGN	/	/	39.52±9.01	30.74±12.82	/	/	/	/	/	/	/	/
FAGCN	/	/	/	/	/	/	36.97±11.25	27.33±7.95	/	/	/	/
DGI	43.82±9.33	40.78±8.06	36.01±7.24	28.11±9.48	46.92±20.14	33.27±9.98	27.29±8.32	21.34±5.51	47.26±7.47	45.23±6.57	44.56±15.54	34.25±10.64
GRAPHPROMPT	**55.51**±10.51	**52.53**±9.76	42.18±9.78	35.12±12.51	48.32±14.02	35.11±6.97	32.73±4.80	26.41±3.88	57.06±9.09	55.81±7.99	**50.55**±17.03	**39.50**±13.23
HECO	/	/	**80.85**±12.79	**80.25**±13.63	/	/	/	/	/	/	/	/
TGN	/	/	/	/	51.28±15.79	36.10±7.96	/	/	/	/	/	/
DDGCL	/	/	/	/	54.12±18.92	36.93±8.88	/	/	/	/	/	/
DSSL	/	/	/	/	/	/	31.34±6.95	24.25±4.12	/	/	/	/
GCOPE	42.56±7.22	42.02±7.33	33.21±1.27	28.50±3.89	25.22±25.68	18.86±14.39	18.41±6.01	15.92±4.16	35.66±6.70	30.32±5.55	41.33±16.67	30.51±7.98
MDGPT	29.41±6.30	28.09±6.01	52.81±9.25	51.01±9.68	43.38±24.31	30.76±12.58	27.53±6.27	22.86±3.96	44.56±8.15	43.01±7.41	44.48±14.06	34.13±7.75
MDGFM	28.15±7.26	26.76±6.98	43.86±11.66	42.93±11.86	**55.73**±30.46	35.46±15.50	23.38±8.39	18.93±6.01	24.76±8.67	23.78±9.23	47.31±21.01	34.24±11.53
SAMGPT	44.16±6.11	42.35±6.21	52.88±9.70	49.94±11.84	54.43±22.11	36.80±10.62	32.81±9.09	26.16±5.34	55.48±8.82	54.41±7.65	47.35±12.75	37.08±8.36
G2P2	45.12±4.84	44.63±4.40	53.75±12.77	49.32±15.61	44.55±14.27	22.31±5.03	34.44±9.05	27.48±5.38	-	-	36.14±6.16	25.53±3.24
GRAPHCLIP	34.46±5.51	32.40±4.87	27.66±5.17	20.17±2.74	31.46±17.08	16.78±7.25	29.70±7.39	25.40±4.97	-	-	39.72±20.55	23.45±8.34
GFT	44.39±6.52	42.65±6.03	45.17±5.90	44.18±5.92	53.12±11.90	**37.33**±5.56	**38.96**±7.47	**32.09**±5.68	40.35±6.01	40.06±5.54	40.51±16.07	32.17±9.12
UNIGRAPH2	46.53±8.79	43.73±7.16	60.82±8.36	57.60±9.45	45.91±23.93	29.62±11.63	25.92±8.21	19.24±4.63	**63.00**±10.09	**62.20**±8.60	44.36±11.05	34.27±5.57

First, similar to the performance on unseen datasets (RQ1), leading GFMs generally outperform supervised approaches, suggesting that broad pre-training can be beneficial to both seen and unseen downstream data. Second, multi-domain knowledge may not always be effectively pre-trained or transferred to target graphs. Specifically, GRAPHPROMPT, which is pre-trained and evaluated on the same graph, remains competitive across many datasets and even outperforms all GFMs on *Cora* and *Elliptic*. This suggests that multi-domain pre-training may favor certain domain-specific knowledge, or that some target graphs may not fully leverage the pre-trained multi-domain representations. In such cases, pre-training directly on the target graph alone can be more effective. Third, GFMs are generally robust across formats. On the dynamic graph *Reddit*, several GFMs (MDGFM, SAMGPT, and GFT) attain better or comparable performance than pre-training methods specifically designed for dynamic graphs (DDGCL and TGN). One possible explaination is that GFMs are trained on the final snapshot, which may preserve structural traces of historical interactions and thereby reduce the need for explicit temporal modeling. Similarly, on the heterophilic graph *Wisconsin*, the heterophily-specific pre-training method DSSL improves over supervised GNNs in Macro-F, but still trails multiple GFMs (GFT, G2P2, SAMGPT, and GRAPHCLIP), whereas FAGCN achieves the strongest format-specific performance but remains comparable to G2P2 and below GFT. These results suggest that GFMs may leverage complementary signals, such as textual attributes in G2P2 or computation-tree representations in GFT, to accommodate heterophilic structures without assuming homophily. A notable exception arises for heterogeneous graphs, such as *ACM*. In particular, HECO, a pre-training method specifically designed for heterogeneous graphs, substantially outperforms all general-purpose GFMs and supervised methods. A potential reason is that heterogeneous graphs contain rich type-specific semantics that may not be fully captured by general pre-training objectives.

Nevertheless, the best GFMs still achieve strong overall performance. More importantly, GFMs offer broader applicability than conventional graph pre-training methods, as pre-trained GFMs can generalize to both seen and unseen downstream graphs across formats, whereas the latter are limited to the seen setting and, in the case of format-aware approaches, to specific formats.

RQ3: *How do GFMs pre-trained on different topic domains perform when adapting to various downstream topics?*

We report one-shot cross-topic adaptation results for **Setting III** in Tables 4 and 28. We further summarize the cross-topic performance and compare with Setting I (broad multi-topic pre-training) by pooling the accuracy of representative GFMs (SAMGPT, G2P2, GFT, and UNIGRAPH2) in Figs. 1 and 2(a). The pooling procedure is described in Appendix D.1.

We first examine whether matching the pre-training and downstream topic domains improves downstream performance. For each downstream topic, Fig. 1 normalizes the accuracy relative to same-topic pre-training, which is set to 100%. Performance is generally strongest when the pre-training and downstream topics match, suggesting a *specialization benefit*. However, this benefit is modest: Setting I with broad multi-topic pre-training, and cross-topic pre-training both remain close to 100% in many cases. A notable exception is the molecular topic, where downstream tasks are graph classification rather than the predominantly node-level tasks in other domains.

Table 4: Accuracy of one-shot node classification on unseen datasets when separately pre-trained on graphs from three different topic domains.

Pre-training topic	Methods	Citation		Social & Web			E-commerce	Finance		Proteins
		Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
Citation (Cora, ACM, ogbn_arxiv, DBLP)	GCOPE	34.12±6.54	0.30±0.04	52.38±27.60	21.17±8.26	22.57±2.53	OOT	33.61±25.24	OOT	48.18±12.68
	MDGPT	38.12±8.03	3.47±0.49	49.47±36.24	27.85±7.76	**28.86**±4.64	15.27±4.77	61.37±25.09	35.75±25.35	**51.76**±11.41
	MDGFM	37.80±5.63	1.10±0.69	49.44±32.82	19.81±7.58	24.55±4.05	6.10±2.15	48.89±21.55	35.70±26.52	50.00±10.10
	SAMGPT	45.77±6.39	**6.38**±0.76	48.48±29.02	24.10±7.05	28.71±5.00	17.84±3.64	54.46±20.93	33.71±18.70	50.47±9.63
	G2P2	45.90±7.69	-	37.81±7.81	28.69±11.51	-	-	60.50±15.97	**37.58**±13.08	-
	GRAPHCLIP	**48.59**±5.92	-	46.21±19.74	**34.22**±5.01	-	-	54.66±23.20	35.47±24.33	-
	GFT	42.97±8.40	2.00±0.22	**62.17**±12.72	26.08±6.11	28.00±4.37	10.53±2.23	53.31±14.23	36.14±13.11	51.18±5.91
	UNIGRAPH2	46.99±11.17	2.76±0.63	45.62±26.98	25.37±9.96	28.09±4.37	**21.33**±4.53	**73.97**±26.92	34.19±14.30	49.95±1.60
Social (Reddit, Texas, Wisconsin, Cornell, IMDB)	GCOPE	36.87±6.91	0.30±0.05	55.14±25.56	21.02±7.78	21.26±3.29	OOT	42.21±21.67	OOT	47.82±11.57
	MDGPT	43.72±7.68	**6.82**±0.73	51.83±25.28	29.29±8.90	**30.12**±4.78	17.35±3.78	55.10±26.93	35.75±11.94	**51.95**±11.74
	MDGFM	38.25±6.34	0.73±0.42	51.72±33.41	18.41±5.24	25.59±3.40	4.06±1.29	57.17±26.96	37.31±23.65	50.11±9.26
	SAMGPT	44.51±9.10	6.55±0.77	49.58±27.96	27.02±7.52	29.34±4.31	**22.30**±4.86	53.38±14.80	36.27±11.56	50.75±11.42
	G2P2	**48.36**±8.67	-	52.52±18.24	**29.64**±11.77	-	-	61.15±16.09	**41.44**±13.88	-
	GRAPHCLIP	37.09±4.30	-	42.38±21.19	22.66±7.26	-	-	53.07±22.98	39.72±22.42	-
	GFT	43.63±8.14	2.08±0.32	**57.92**±27.98	28.70±6.04	28.40±4.24	10.38±2.14	60.38±20.04	35.60±13.75	50.97±5.91
	UNIGRAPH2	47.07±10.67	2.46±0.64	44.25±28.39	26.36±9.66	28.42±3.77	19.12±4.79	**81.04**±29.69	35.74±15.97	50.07±3.76
Molecule (HIV, COX2)	GCOPE	32.06±7.78	0.23±0.04	31.83±34.52	20.24±8.86	20.92±1.83	OOT	45.96±19.43	OOT	47.41±13.79
	MDGPT	46.39±8.34	**6.39**±0.84	48.96±16.14	**29.59**±8.15	29.31±5.05	24.01±6.18	59.06±30.68	35.40±11.50	**51.97**±11.69
	MDGFM	39.82±7.70	0.99±0.40	51.82±31.05	19.90±7.21	25.12±4.70	9.27±4.32	51.35±23.42	39.27±17.63	49.14±9.40
	SAMGPT	**48.60**±7.03	6.19±0.69	51.29±27.91	25.71±8.36	**29.48**±4.46	**24.57**±6.17	55.24±16.84	40.15±16.93	51.91±10.10
	G2P2	44.83±7.61	-	52.71±20.18	28.79±11.05	-	-	62.35±16.39	**41.46**±12.19	-
	GRAPHCLIP	34.57±3.58	-	33.16±21.28	21.80±7.11	-	-	50.59±23.43	38.22±26.78	-
	GFT	44.64±9.22	2.03±0.26	**60.07**±27.64	27.41±6.62	28.34±4.05	10.62±2.12	65.37±23.42	36.35±14.15	51.07±6.21
	UNIGRAPH2	45.27±10.44	1.81±0.49	43.88±25.10	23.93±8.05	28.16±3.62	11.21±2.27	**77.43**±26.08	36.83±13.58	50.00±4.08

Table 5: Evaluation of one-shot node classification on format domain adaptation.

Methods	Heterogeneous ogbn-mag		Heterophilic Chameleon		Heterophilic Actor		Dynamic Wikipedia		Textual Pubmed		Textual Products	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE	0.32±0.04	0.11±0.01	22.46±2.93	19.99±3.38	20.66±8.01	16.05±4.58	**58.69**±29.29	**36.89**±14.91	39.06±7.55	35.09±8.07	OOT	OOT
MDGPT	6.01±0.81	4.42±0.37	27.90±5.13	**25.06**±5.03	**27.23**±8.64	**22.22**±5.09	54.17±22.13	35.72±10.71	45.80±10.14	45.10±10.47	21.23±4.96	11.50±1.88
MDGFM	0.92±0.54	0.56±0.41	21.50±3.12	17.54±4.00	20.20±7.73	13.83±3.12	50.66±29.90	32.51±14.61	34.51±5.84	28.24±4.55	3.92±2.17	1.78±0.47
SAMGPT	**7.35**±0.73	**5.49**±0.27	**28.14**±4.77	24.48±4.66	24.58±6.49	19.21±4.17	55.95±31.62	34.51±15.82	**50.00**±7.93	**48.68**±8.50	**26.10**±5.55	**13.38**±1.52
G2P2	-	-	-	-	27.04±10.41	19.83±6.23	54.78±16.60	28.69±8.64	44.78±7.92	41.63±7.81	-	-
GRAPHCLIP	-	-	-	-	22.82±5.34	16.50±2.44	43.23±20.45	20.90±7.66	33.68±4.22	30.95±3.54	-	-
GFT	1.71±1.48	0.54±0.33	26.31±4.17	22.15±4.35	24.29±7.76	19.22±4.37	56.94±31.46	36.33±16.66	45.32±7.89	43.17±8.21	6.30±4.34	3.26±1.18
UNIGRAPH2	2.36±1.32	0.66±0.78	25.96±4.36	23.77±4.31	22.52±8.12	17.12±4.23	54.87±24.50	35.63±11.55	45.97±9.20	44.60±9.27	12.91±8.53	7.38±1.05

Without explicit graph-level pre-training objectives or readout adaptation, the benefit of topic matching may be outweighed by the scale and diversity of broad multi-topic pre-training. We next assess robustness against broader topics. Figure 2(a) reports, for each downstream topic domain, the fraction of models in which single-topic pre-training in Setting III outperforms multi-topic pre-training in Setting I. The results show that topic-specific pre-training can be competitive when the downstream topic is the same or closely related. For example, pre-training exclusively on social graphs outperforms broad multi-topic pre-training more frequently on citation, social, and finance downstream topics, which are all related to human activities. However, its overall overperformance rate is not dominant, and the benefit does not generalize to other topic domains.

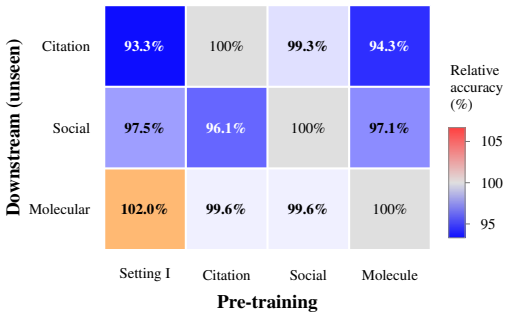
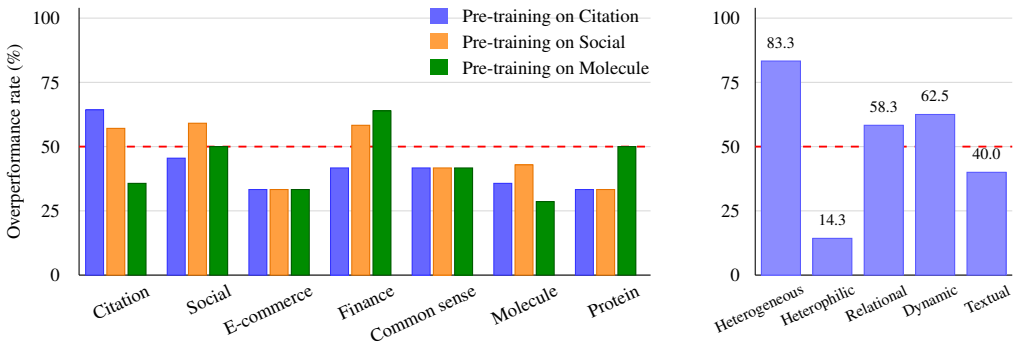


Figure 1: Pooled accuracy across pre-training and downstream topic domains, with each row normalized by the result obtained when the pre-training and downstream topics match.

In summary, the meta-analysis based on Figs. 1 and 2(a) shows that pre-training on the same or similar topics can yield a slight advantage, whereas the broad pre-training in Setting I delivers the most stable performance across the full range of downstream topics.

RQ4: *How do GFMs pre-trained on a base format domain differ from those pre-trained on a multi-format domain when adapting to downstream format domains?* We report one-shot cross-format performance for **Setting IV** in Tables 5 and 33, with additional comparisons against format-specific backbones in Table 6. We further compare Setting IV with Setting I in Fig. 2(b), following the same meta-analysis setup as in RQ3.



(a) Setting III vs. Setting I (b) Setting IV vs. Setting I

Figure 2: Pooled overperformance rates relative to Setting I for (a) cross-topic adaptation in Setting III, and (b) cross-format adaptation in Setting IV. Bars below 50% favor Setting I.

Table 6: Effects of the heterophily-aware backbone (FAGCN) on *Chameleon* and *Actor*. Δ (%) refers to the relative change when switching the homophilic backbone to FAGCN.

Methods	Chameleon						Actor					
	homophilic backbone		FAGCN backbone		Δ (%)		homophilic backbone		FAGCN backbone		Δ (%)	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE	22.32±1.86	18.99±2.51	22.46±2.93	19.99±3.38	+0.63%	+5.27%	20.07±7.27	14.33±3.34	20.66±8.01	16.05±4.58	+2.94%	+12.00%
MDGPT	27.90±5.13	**25.06**±5.03	27.62±4.41	23.83±4.36	-1.00%	-4.91%	**27.23**±8.64	**22.22**±5.09	25.85±6.33	20.59±4.38	-5.07%	-7.34%
MDGFM	21.50±3.12	17.54±4.00	26.55±4.10	24.75±3.59	+23.49%	+41.11%	20.20±7.73	13.83±3.12	**26.26**±4.22	**21.89**±3.25	+30.00%	+58.28%
SAMGPT	**28.14**±4.77	24.48±4.66	29.21±4.39	25.26±4.18	+3.80%	+3.19%	24.58±6.49	19.21±4.17	25.28±8.50	18.56±4.38	+2.85%	-3.38%
G2P2	-	-	-	-	-	-	27.04±10.41	19.83±6.23	21.36±10.99	13.18±4.70	-21.01%	-33.54%
GRAPHCLIP	-	-	-	-	-	-	22.82±5.34	16.50±2.44	19.22±18.62	6.76±5.42	-15.78%	-59.03%
GFT	26.31±4.17	22.15±4.35	25.89±3.09	20.11±3.92	-1.60%	-9.21%	24.29±7.76	19.22±4.37	24.71±6.12	19.29±3.24	+1.73%	+0.36%
UNIGRAPH2	25.96±4.36	23.77±4.31	**31.05**±4.54	**26.88**±4.68	+19.61%	+13.08%	22.52±8.12	17.12±4.23	25.08±7.67	19.11±4.57	+11.37%	+11.62%

For the homophilic backbone, all GFMs except GCOPE retain their original, homophilic backbones; for the heterophilic backbone, all GFMs use FAGCN. Since GCOPE already uses FAGCN in its original implementation, we replace it with GCN for its homophilic-backbone variant.

First, the effectiveness of base-format pre-training varies substantially across formats. As shown in Fig. 2(b), the overperformance rates of Setting IV are 83.3%, 14.3%, 58.3%, 62.5%, and 40.0% on heterogeneous, heterophilic, relational, dynamic, and textual formats, respectively. The particularly low rate on heterophilic graphs suggests a direct mismatch between the homophilic base-format corpus and heterophilic targets: pre-training without disassortative structures may reinforce homophily-oriented representations that are less transferable to downstream heterophilic graphs. Second, when adapting to text-attributed graphs, GFMs with text encoders (e.g., G2P2, GRAPHCLIP, GFT, UNIGRAPH2) pre-trained on the base format (Setting IV) generally tend to underperform pre-training with multiple formats (Setting I). The reason is that, under Setting IV, textual information is absent during pre-training, leaving their text encoders with little meaningful supervision. When text becomes available downstream, the additional modality behaves more like noise than a useful signal, leading to weaker transfer. Additionally, Table 6 examines whether a heterophily-aware backbone such as FAGCN can improve GFM generalization to heterophilic downstream graphs such as *Chameleon* and *Actor*. The results suggest that the FAGCN backbone can change the ranking across GFMs, but it does not uniformly solve format transfer by itself. Specifically, relative to their homophilic backbone, FAGCN substantially improves MDGFM and benefits UNIGRAPH2 on both datasets, while the gains for other GFMs are mixed or negative. Overall, format transfer cannot be attributed to backbone design alone, but also depends on how the backbone interacts with the pre-training objectives and the learned representations.

5.2 DISCUSSION OF INSIGHTS AND FUTURE DIRECTIONS

Settings I and II: Adapting to unseen and seen datasets. First, *GFMs provide useful but uneven gains on both unseen and seen datasets, indicating that multi-domain knowledge integration remains a central bottleneck.* On unseen datasets, the strongest GFMs often improve over supervised GNNs, but no method dominates consistently. On seen datasets, single-domain target-specific pre-training can occasionally outperform broad multi-domain pre-training. Together, these observations reveal two related limitations: insufficient transfer to unseen targets and interference among domains

within the pre-training mixture. *Future work* should therefore improve target-aware adaptation and balance domain contributions through sampling, adaptive weighting, or mixture-based routing.

Second, *broad pre-training supports cross-format generalization, but reliable format awareness remains incomplete.* The strongest GFMs can match or exceed single-domain format-specific pre-training methods on dynamic and heterophilic graphs, implying that transferable mechanisms such as GFT's computation-tree vocabulary can capture structural cues beyond the backbone. Heterogeneous graphs remain a notable exception and favor explicit type-rich modeling, motivating *future work* on format-aware objectives and modular components rather than a single uniform objective.

Setting III: Adapting across topic domains. First, *while pre-training on graphs from same or similar topics provides useful domain-specific priors, broader topic coverage often strengthens GFMs and yields more robust adaptation across downstream domains.* In particular, the specialization benefit of topic alignment between pre-training and downstream graphs is often modest and does not consistently manifest across all topics. In contrast, broad multi-topic pre-training remains competitive across various downstream topics and is less prone to severe cross-topic degradation. This points to a concrete direction for building *future GFMs*: expanding and balancing topic domains in pre-training to promote broader cross-topic generalization while mitigating negative transfer.

Second, *coarse topic alignment is not a sufficient indicator of the specialization benefit.* Coarse topic categories can mask substantial dataset-level differences, including discrepancies in degree distributions and textual feature similarity, as illustrated in Fig. 5 and further elaborated in Appendix C.4. These finer-grained differences can dominate transfer outcomes even within the same nominal topic. Additionally, the results on the molecular topic further reveal a task-granularity gap between graph classification and the predominantly node-level pre-training objectives; without explicit graph-level objectives or readout adaptation, current GFMs may not fully exploit molecule-domain priors. Accordingly, *future design* of pre-training strategies should account for dataset-level diversity as well as the granularity of pre-training objectives, beyond assuming coarse topic proximity.

Setting IV: Format domain adaptation. *GFMs show encouraging potential for cross-format adaptation, but their reliability depends on the nature of the gap between the target and base formats.* Specifically, pre-training on the base format alone remains competitive when transferring to heterogeneous, relational, and dynamic graphs, but underperforms broad multi-format pre-training on heterophilic and text-attributed graphs. One plausible explanation is pre-training exposure: heterophily and textual attributes alter core assumptions of the base setting and are absent from the homophilic, text-free pre-training corpus, whereas heterogeneity, multiple relations, and temporal dynamics can be viewed as additional structure layered on top of the base graph. Broad multi-format pre-training therefore provides direct exposure to heterophilic and textual signals that the base-format corpus lacks. These observations point to several *promising directions* for future GFMs. First, pre-training corpora could be designed to span meaningful variation along each format axis, rather than merely include isolated examples. For instance, covering a range of homophily levels may encourage aggregation mechanisms that adapt to heterophily rather than rely on the homophily assumption, while textual information could be incorporated as an optional enhancement built on a robust text-free pathway. Second, the limited gains from broader format mixing on heterogeneous, relational, and dynamic targets suggest that future models may benefit from explicitly type-, relation-, and time-aware objectives or architectures, rather than relying on exposure alone.

6 Conclusions

We introduced a new benchmark for GFMs that treats domain shift as inherently two-dimensional, disentangling topic semantics from graph formats. We comprehensively evaluated eight recent GFMs on 33 datasets spanning seven topic domains and six format domains under four complementary settings: (i) pre-training on diverse topics and formats and adapting to unseen graphs; (ii) adapting the same pre-trained models to seen graphs; (iii) pre-training on a single topic domain and adapting across topics; and (iv) pre-training on a canonical base format and adapting across formats. Across these settings, our results show that current GFMs exhibit promising transfer capabilities, while there remains room to improve their robustness across domains. Finally, We discussed key insights from the benchmark and highlighted several directions for improving the generalization and robustness of future GFMs.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data for one of the evaluated baselines: following the original GRAPHCLIP protocol, we used OpenAI's GPT-5-MINI to summarize sampled ego-subgraphs into the natural language descriptions that serve as its pre-training text input, as described in Appendix B.3. We note separately that the *Cora* and *Pubmed* label descriptions we adopted were generated with GPT-3.5-TURBO by their original providers (Liu et al., 2024), and were used by us as released.

Additionally, we used generative AI tools solely for editing the paper to improve readability. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design our benchmark methodology or evaluation protocol, or interpret our experimental results. All technical content, claims, and conclusions are our own. We have reviewed all AI-assisted work and take full responsibility for the contents of this submission.

ETHICS STATEMENT

This work is a benchmark study built entirely on publicly available graph datasets, and introduces no new data collection involving human subjects. Two of the datasets we evaluate, *DGraph* and *Elliptic*, are derived from real-world financial activities. Both are released by their original providers in anonymized form: *DGraph* node features are anonymized user attributes, and *Elliptic* transactions are identified only by hashed identifiers. We use both strictly under their original licenses, and we make no attempt to re-identify any individual or entity. We do not foresee any other ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide our full codebase, data processing pipeline, and evaluation scripts at `https://anonymous.4open.science/r/GFMBenchmark-7EA0`. The unified data object interface and all dataset pre-processing protocols, including the homogenization of heterogeneous graphs, temporal snapshot construction, and cross-dataset feature-dimension alignment, are specified in Appendix B.1. The few-shot task construction and model-specific sampling protocols are described in Appendix B.2, and the complete hardware environment, per-model backbones, and hyperparameter configurations are reported in Appendix B.3. Every number we report is the mean and standard deviation over 50 few-shot episodes evaluated with five random seeds each, i.e., 250 runs per dataset and shot setting.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Marco Bardoscia, Paolo Barucca, Stefano Battiston, Fabio Caccioli, Giulio Cimini, Diego Garlaschelli, Fabio Saracco, Tiziano Squartini, and Guido Caldarelli. The physics of financial networks. *Nature Reviews Physics*, 3(7):490–507, 2021.

Wendong Bi, Lun Du, Qiang Fu, Yanlin Wang, Shi Han, and Dongmei Zhang. Make heterophilic graphs better fit gnn: A graph rewiring approach. *TKDE*, 2024.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3950–3957, 2021.

Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka, and Tom Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, pp. 1306–1313, 2010.

Yukuo Cen, Xu Zou, Jianwei Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. Representation learning for attributed multiplex heterogeneous network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1358–1368, 2019.

Runjin Chen, Tong Zhao, Ajay Kumar Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. In *ICML*, pp. 7809–7823, 2024a.

Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Appl.*, 141, 2020.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. Exploring the potential of large language models (llms) in learning on graphs. *SIGKDD Explor.*, 25(2):42–61, 2023.

Zhikai Chen, Haitao Mao, Jingzhe Liu, Yu Song, Bingheng Li, Wei Jin, Bahare Fatemi, Anton Tsitsulin, Bryan Perozzi, Hui Liu, et al. Text-space graph foundation models: Comprehensive benchmarks and new insights. In *NeurIPS*, volume 37, pp. 7464–7492, 2024b.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, pp. 135–144, 2017.

Dmitry Eremeev, Gleb Bazhenov, Oleg Platonov, Artem Babenko, and Liudmila Prokhorenkova. Turning tabular foundation models into graph foundation models. *arXiv preprint arXiv:2508.20906*, 2025.

Dmitry Eremeev, Oleg Platonov, Gleb Bazhenov, Artem Babenko, and Liudmila Prokhorenkova. GraphPFN: A prior-data fitted graph foundation model. In *ICML*, 2026.

Linton Freeman et al. The development of social network analysis. *A Study in the Sociology of Science*, 1(687):159–167, 2004.

Alessandro Giuliani, Arun Krishnan, Joseph P Zbilut, and Masaru Tomita. Proteins as networks: usefulness of graph theory in protein science. *Current Protein and Peptide Science*, 9(1):28–38, 2008.

Andrew G Haldane. Rethinking the financial network. In *Fragile stabilität–stabile fragilität*, pp. 243–278. Springer, 2013.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, volume 30, 2017.

Frank Harary and Gopal Gupta. Dynamic graph models. *Mathematical and Computer Modelling*, 25(7):79–87, 1997.

Harrison Hartle, Brennan Klein, Stefan McCabe, Alexander Daniels, Guillaume St-Onge, Charles Murphy, and Laurent Hébert-Dufresne. Network comparison and the within-ensemble graph distance. *Proceedings. Mathematical, Physical, and Engineering Sciences*, 476(2243):20190744, 2020.

Adrian Hayler, Xingyue Huang, Ismail Ilkan Ceylan, Michael Bronstein, and Ben Finkelshtein. Bringing graphs to the table: Zero-shot node classification via tabular foundation models. *arXiv preprint arXiv:2509.07143*, 2025.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *WWW*, pp. 1759–1770, 2025.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, pp. 22118–22133, 2020a.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *SIGKDD*, pp. 1857–1867, 2020b.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy Liang, and Jure Leskovec. Prodigy: Enabling in-context learning over graphs. In *NeurIPS*, 2023.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. *Advances in Neural Information Processing Systems*, 35:22765–22777, 2022.

Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608, 2016.

Jinwoo Kim, Xingyue Huang, Krzysztof Olejniczak, Kyungbin Min, Michael Bronstein, Seunghoon Hong, and Ismail I Ceylan. Flock: A knowledge graph foundation model via learning on random walks. In *ICLR*, volume 2026, pp. 36664–36705, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1269–1278, 2019.

Shujie Li, Yuxia Wu, Yuan Fang, and Chuan Shi. Hetgb: A comprehensive benchmark for heterophilic text-attributed graphs. *arXiv preprint arXiv:2503.04822*, 2025.

Yuhan Li, Peisong Wang, Zhixun Li, Jeffrey Xu Yu, and Jia Li. Zerog: Investigating cross-dataset zero-shot transferability in graphs. In *SIGKDD*, pp. 1725–1735, 2024.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *ICLR*, 2024.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. Graph foundation models: Concepts, opportunities and challenges. *IEEE TPAMI*, 2025a.

Shikun Liu, Deyu Zou, Nima Shoghi, Victor Fung, Kai Liu, and Pan Li. Roft-mol: Benchmarking robust fine-tuning with molecular graph foundation models. In *NeurIPS*, 2025b.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *WWW*, pp. 417–428, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress?: Revisiting, benchmarking and refining heterogeneous graph neural networks. In *KDD*, pp. 1150–1160, 2021.

Franco Manessi, Alessandro Rozza, and Mario Manzo. Dynamic graph convolutional networks. *Pattern Recognition*, 97:107000, 2020.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *ICML*, 2024.

Alexandra Marin and Barry Wellman. Social network analysis: An introduction. *The SAGE handbook of social network analysis*, 11:25, 2011.

Colin D McLaren and Mark W Bruner. Citation network analysis. *International Review of Sport and Exercise Psychology*, 15(1):179–198, 2022.

Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

Oleg Platonov, Gleb Bazhenov, Dmitry Eremeev, and Liudmila Prokhorenkova. A fair evaluation of graph foundation models for node property prediction. In *Workshop on Graph Foundation Models at ICML*, 2026.

Filippo Radicchi, Santo Fortunato, and Alessandro Vespignani. Citation networks. *Models of science dynamics: Encounters between complexity theory and information sciences*, pp. 233–257, 2011.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *NeurIPS*, pp. 14501–14515, 2022.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on Graph Representation Learning*, 2020.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M Bronstein. Edge directionality improves learning on heterophilic graphs. In *Learning on graphs conference*, pp. 25–1, 2024.

Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Junhua Shi, Qingyun Sun, Haonan Yuan, and Xingcheng Fu. SA^{2}GFM: Enhancing robust graph foundation models with structure-aware semantic augmentation. In *AAAI*, volume 40, pp. 15716–15724, 2026.

Li Sun, Zhenhao Huang, Suyang Zhou, Qiqi Wan, Hao Peng, and Philip Yu. RiemannGFM: Learning a graph foundation model from Riemannian geometry. In *WWW*, pp. 1154–1165, 2025.

Li Sun, Zhenhao Huang, Silei Chen, Lanxu Yang, Junda Ye, Sen Su, and Philip S. Yu. Multi-domain Riemannian graph gluing for building graph foundation models. In *ICLR*, 2026.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *SIGIR*, pp. 491–500, 2024a.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. HiGPT: Heterogeneous graph language model. In *KDD*, pp. 2842–2853, 2024b.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pp. 21076–21089. PMLR, 2022.

Sheng Tian, Ruofan Wu, Leilei Shi, Liang Zhu, and Tao Xiong. Self-supervised representation learning on dynamic graphs. In *CIKM*, pp. 1814–1823, 2021.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pp. 57–66, 2015.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018a.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR*, 2018b.

Shuo Wang, Bokui Wang, Zhixiang Shen, Boyan Deng, and Zhao Kang. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. In *ICML*, 2025.

Shuo Wang, Xiangyu Wang, Quanxin Wang, Bailin Wu, Bokui Wang, Shunyang Huang, Boyan Deng, Haonan Liu, Ruiyi Fang, Zhenxiang Xu, et al. The post-gcn decade revisited: Curvature-stratified evaluation of relational learning. *arXiv preprint arXiv:2606.06397*, 2026.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *WWW*, pp. 2022–2032, 2019.

Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *KDD*, pp. 1726–1736, 2021a.

Xiao Wang, Deyu Bo, Chuan Shi, Shaohua Fan, Yanfang Ye, and Philip S Yu. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. *TBG*, 9(2):415–436, 2022.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021b.

Zehong Wang, Zheyuan Zhang, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Gft: Graph foundation model with transferable tree vocabulary. In *NeurIPS*, volume 37, pp. 107403–107443, 2024.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019.

Zhihao Wen and Yuan Fang. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *SIGIR*, pp. 506–516, 2023.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *ICML*, 2025.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Lianghao Xia, Ben Kao, and Chao Huang. Opengraph: Towards open graph foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2365–2379, 2024.

Teng Xiao, Zhengyu Chen, Zhimeng Guo, Zeyang Zhuang, and Suhang Wang. Decoupled self-supervised learning for graphs. In *NeurIPS*, volume 35, pp. 620–634, 2022.

Yuhao Xu, Xinqi Liu, Keyu Duan, Yi Fang, Yu-Neng Chuang, Daochen Zha, and Qiaoyu Tan. Graphfm: A comprehensive benchmark for graph foundation model. *arXiv preprint arXiv:2406.08310*, 2024.

Jinyu Yang, Liangwei Yang, Zeyuan Guo, Jiayi Gao, Jing Wu, Tianhao Chai, Hai Huang, Cheng Yang, and Chuan Shi. Benchmarking graph foundation models. In *SIGKDD*, pp. 5866–5875, 2025.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *NeurIPS*, volume 31, 2018.

Xingtong Yu, Zhenghao Liu, Yuan Fang, Zemin Liu, Sihong Chen, and Xinming Zhang. Generalized graph prompt: Toward a unification of pre-training and downstream tasks on graphs. *IEEE TKDE*, 2024a.

Xingtong Yu, Chang Zhou, Yuan Fang, and Xinming Zhang. Text-free multi-domain graph pre-training: Toward graph foundation models. *arXiv preprint arXiv:2405.13934*, 2024b.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *WWW*, pp. 1142–1153, 2025a.

Xingtong Yu, Zhenghao Liu, Yuan Fang, and Xinming Zhang. Dygprompt: Learning feature and time prompts on dynamic graphs. In *ICLR*, 2025b.

Xingtong Yu, Jie Zhang, Yuan Fang, and Renhe Jiang. Non-homophilic graph pre-training and prompt learning. In *SIGKDD*, pp. 1844–1854, 2025c.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S Yu. Graver: Generative graph vocabularies for robust graph foundation models fine-tuning. In *NeurIPS*, 2025.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *SIGKDD*, pp. 793–803, 2019.

Haihong Zhao, Aochuan Chen, Xiangguo Sun, Hong Cheng, and Jia Li. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *KDD*, pp. 4443–4454, 2024.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. In *ICLR*, 2023.

Jianan Zhao, Zhaocheng Zhu, Mikhail Galkin, Hesham Mostafa, Michael Bronstein, and Jian Tang. Fully-inductive node classification on arbitrary graphs. In *ICLR*, volume 2025, pp. 23968–23986, 2025.

Qi Zhao, Ze Ye, Chao Chen, and Yusu Wang. Persistence enhanced graph neural network. In *ICAIS*, pp. 2896–2906, 2020.

Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *WWW*, pp. 2183–2197, 2025.

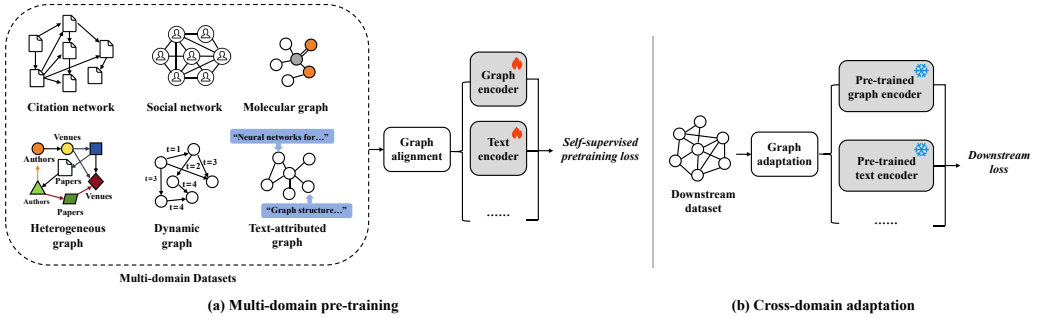


Figure 3: Graph foundation models pipeline.

A Benchmark Components

A.1 GFM Pipleline

We focus on GFMs that follow the pipeline in Fig. 3. These GFMs typically operates in two stages: (i) self-supervised pre-training on a broad multi-domain corpus of graphs, and (ii) cross-domain few-shot adaptation to various downstream graphs

Specifically, let $\mathcal{G}$ denote a collection of graphs from various domains. Given a graph $G \in \mathcal{G}$ with N nodes, a GFM specifies an encoder $\texttt{GE}(\cdot;\theta) : \mathcal{G} \to \mathbb{R}^{N\times d}$ that maps G to d-dimensional node representations[1]. In the pre-training stage, given an *unlabeled* multi-domain graph corpus $\mathcal{D}_{\text{pre}} \subset \mathcal{G}$, the encoder's parameters are learned by minimizing a self-supervised objective $\mathcal{L}_{\text{pre}}$:

$$\theta_{\text{pre}} = \arg\min_{\theta} \ \mathbb{E}_{G\in\mathcal{D}_{\text{pre}}} \mathcal{L}_{\text{pre}}\big(\texttt{GE}(G;\theta)\big). \tag{1}$$

In the downstream stage, assume task-specific data $(\mathcal{D}_{\text{down}}, \mathcal{Y}_{\text{down}})$, where $\mathcal{D}_{\text{down}} \subset \mathcal{G}$ denotes the task graphs, and $\mathcal{Y}_{\text{down}}$ denotes labels associated with the task. In few-shot settings, $\mathcal{Y}_{\text{down}}$ contains only K labeled instances per class for some small K. A lightweight task adapter f_ϕ is optimized using a supervised loss $\mathcal{L}_{\text{down}}$, while the pre-trained weights θ_{pre} are frozen, as follows:

$$\min_{\phi} \ \mathbb{E}_{(G,Y)\in(\mathcal{D}_{\text{down}},\mathcal{Y}_{\text{down}})} \mathcal{L}_{\text{down}}\Big(f_\phi\big(\texttt{GE}(G;\theta_{\text{pre}})\big), Y\Big). \tag{2}$$

A.2 Benchmark Pipeline

We illustrate the four evaluation settings in Fig. 4.

A.3 Evaluated and GFM-style Methods

We summarize the representative GFMs in Table 7. The column *Labeled source* indicates whether a method relies on supervision, such as labeled objectives or instruction-tuning signals, during pre-training. The *Alignment* columns categorize the primary mechanisms used to reduce domain gaps: *Structure* and *Feature* denote whether the method explicitly aligns cross-domain structural patterns or feature spaces, respectively; *Codebook* indicates whether a learnable codebook is introduced for discretization or alignment; *Geometry* indicates whether non-Euclidean (Riemannian) geometry is used to align domains; *Graph mod.* and *Feature mod.* indicate whether the method modifies the input graph topology or transforms features to facilitate alignment; and *Textual* indicates whether textual information is used to bridge domain gaps. The *PLM/LLM* column indicates whether a pretrained language model is a core component (e.g., for feature extraction, cross-modal alignment, prediction, or graph/text generation). *Adaptation* specifies the primary downstream adaptation strategy, such as fine-tuning or prompting. *Zero-shot* column indicates whether the method can classify unseen downstream datasets without labeled target samples.

[1] Minimal notation is used for brevity.

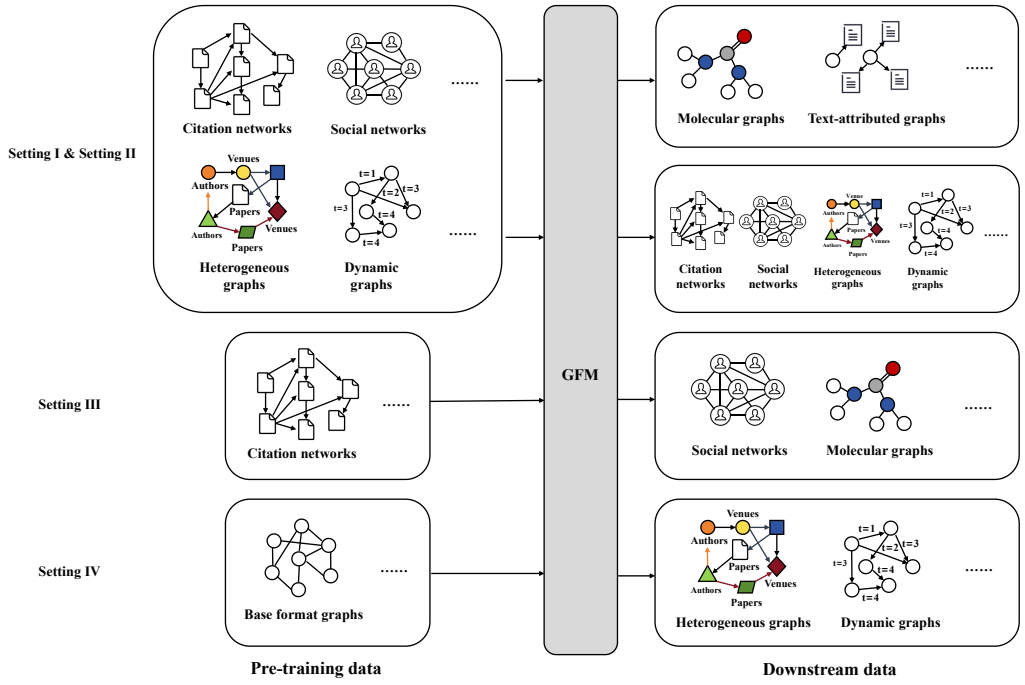


Figure 4: Four evaluation settings in our benchmark.

Table 7: Comparison of representative GFMs.

Methods	Labeled	Alignment						PLM/LLM	Adapt.	Zero-shot
		Struct.	Feat.	Code	Graph mod.	Feat. mod.	Text			
Benchmarked graph foundation models										
GCOPE (Zhao et al., 2024)	×	×	✓	×	✓	×	×	×	F, P	×
MDGPT (Yu et al., 2024b)	×	×	✓	×	×	✓	×	×	P	×
MDGFM (Wang et al., 2025)	×	×	✓	×	✓	✓	×	×	P	×
SAMGPT (Yu et al., 2025a)	×	✓	✓	×	×	✓	×	×	P	×
G2P2 (Wen & Fang, 2023)	×	×	✓	×	×	×	✓	E, A	P	✓
GRAPHCLIP (Zhu et al., 2025)	×	×	✓	×	×	×	✓	E, A	P	✓
GFT (Wang et al., 2024)	×	✓	✓	✓	×	×	✓	E	F, P	×
UNIGRAPH2 (He et al., 2025)	×	✓	✓	×	×	✓	✓	E	*I*	×
Other GFM-style methods requiring labeled source data (out of scope)										
OFA (Liu et al., 2024)	✓	×	✓	×	✓	×	✓	E	P	✓
PRODIGY (Huang et al., 2023)	✓	×	✓	×	✓	×	×	×	P	×
GRAPHGPT (Tang et al., 2024a)	✓	×	✓	×	×	×	✓	A, P	P	✓
HIGPT (Tang et al., 2024b)	✓	×	✓	×	×	×	✓	A, P	P	✓
OPENGRAPH (Xia et al., 2024)	✓	✓	✓	×	✓	×	×	G	×	✓
GRAVER (Yuan et al., 2025)	✓	✓	×	×	✓	✓	×	E	F, P	✓
LLAGA (Chen et al., 2024a)	✓	×	✓	×	×	×	✓	P	P	✓
ZEROG (Li et al., 2024)	✓	×	✓	×	×	×	✓	A	P	✓
GRAPHANY (Zhao et al., 2025)	✓	✓	✓	×	×	✓	×	×	F	×

Under *PLM/LLM*, 'E': extractor, 'A': alignment, 'P': predictor, and 'G': generator. Under *Adaptation*, 'F': fine-tuning, 'P': prompting, and 'I': in-context learning.

Evaluated GFMs. We evaluate a set of representative GFMs that embody different mechanism designs for mitigating domain gaps. Concretely, we include:

- GCOPE (Zhao et al., 2024) builds a joint training graph by adding a learnable coordinator (virtual) node for each domain and connecting it to all nodes in that domain and to other coordinators to align various domains.
- MDGPT (Yu et al., 2024b) introduces domain tokens to harmonize heterogeneous feature spaces, and adopts a dual-prompt design to condition the encoder on domain-specific characteristics during adaptation.

- SAMGPT (Yu et al., 2025a) extends MDGPT by explicitly addressing structural discrepancies across domains. It injects structure-aware tokens and prompts to modulate the message-passing process in the graph encoder, thereby unifying domain-dependent structural inductive biases.
- MDGFM (Wang et al., 2025) refines the topology of each source and target graph via graph structure learning, where node features and aggregated neighbor features are fused by a balance token and sparsified with kNN, and adapts to downstream domains with dual prompts, mitigating structural noise and discrepancies for cross-domain transfer.
- G2P2 (Wen & Fang, 2023) jointly pre-trains a text encoder and a graph encoder via graph-grounded contrastive learning. It leverages a document-node bijection and neighborhood-derived summary text to construct positive pairs, and optimizes multiple interaction losses.
- GRAPHCLIP (Zhu et al., 2025) employs a LLM to generate natural language summaries for sampled subgraphs, then pre-trains a graph encoder with a CLIP-style contrastive objective that aligns graph embeddings with summary text embeddings.
- GFT (Wang et al., 2024) learns a transferable tree vocabulary by treating the message-passing computation trees induced by a GNN as discrete structural tokens, and pretrains a single backbone with a computation-tree reconstruction objective so it can transfer robustly across domains and tasks while mitigating negative transfer.
- UNIGRAPH2 (He et al., 2025) learns a unified embedding space for multi-modal graphs by binding heterogeneous modalities into a shared representation, enabling consistent cross-modal semantics and structure-aware fusion.

Other representative GFM-style methods. We further discuss other GFM-style methods. Note that these methods are excluded from our primary evaluation as they typically require labels in the pre-training datasets, which diverges from our experimental setting. Concretely, we discuss:

- OFA (Liu et al., 2024) unifies diverse graph learning tasks into a single framework by representing all data as TAGs and introducing nodes-of-interest prompts to standardize task specifications across domains.
- PRODIGY (Huang et al., 2023) introduces an in-context learning framework that unifies diverse tasks using a prompt graph representation, and optimizes a self-supervised neighbor matching objective to facilitate out-of-the-box generalization.
- GRAPHGPT (Tang et al., 2024a) aligns graph structural knowledge with LLMs via a dual-stage instruction tuning paradigm, utilizing a self-supervised graph matching task for structural grounding and Chain-of-Thought distillation to enhance step-by-step reasoning and generalization capabilities.
- HIGPT (Tang et al., 2024b) extends GraphGPT by explicitly addressing relation heterogeneity and distribution shifts across domains. It introduces an in-context heterogeneous graph tokenizer and Mixture-of-Thought instructions to align diverse node and edge semantics, thereby unifying domain-dependent heterogeneity patterns for zero-shot adaptation.
- OPENGRAPH (Xia et al., 2024) develops a scalable zero-shot foundation model that employs a unified topological tokenizer and graph transformer to resolve node shifts and capture dependencies, while utilizing LLM-based augmentation to mitigate data scarcity.
- GRAVER (Yuan et al., 2025) stabilizes few-shot fine-tuning by augmenting support samples with adaptive graph vocabularies generated via graphon-based experts and a mixture-of-experts routing network.
- LLAGA (Chen et al., 2024a) adapts frozen LLMs by organizing node neighborhoods into structure-preserving sequences via parameter-free templates and mapping them to the token space using a versatile projector trained on unified QA tasks.
- ZEROG (Li et al., 2024) facilitates cross-dataset zero-shot transfer by mapping node attributes and class descriptions into a unified semantic space via a pre-trained LM, utilizing prompt-based subgraph sampling to mitigate negative transfer and reformulating classification as a text similarity task.
- GRAPHANY (Zhao et al., 2025) establishes a fully inductive foundation model for node classification on arbitrary graphs by deriving analytical solutions via LinearGNNs and fusing them using an attention module based on entropy-normalized distance features.

Supervised learning methods. We evaluate representative supervised GNNs as task-specific baselines, tailored for homogeneous, heterogeneous, and heterophilic graphs:

- GCN (Kipf & Welling, 2017) employs a normalized neighborhood aggregation that effectively acts as mean-pooling over adjacent nodes' features, enabling integration of local structural information under the homophily assumption.
- GAT (Veličković et al., 2018a) enhances neighborhood aggregation by incorporating a learnable attention mechanism that assigns distinct importance weights to each neighbor, enabling the model to emphasize more relevant connections during message passing adaptively.
- SIMPLE-HGN (Lv et al., 2021) extends attention-based aggregation to heterogeneous graphs by incorporating edge type embeddings into the attention computation, allowing relation-aware inter-type message passing.
- FAGCN (Bo et al., 2021) introduces a frequency adaptation mechanism with a self-gating filter that adaptively integrates both low-frequency (smooth) and high-frequency (discriminative) signals in propagation, thereby enhancing robustness in heterophilic settings.

Single-domain pre-training and adaptation methods. We further compare against state-of-the-art self-supervised pre-training methods. They are pre-trained on a single domain and adapt to the same domain. These methods are categorized by their adaptation paradigms, covering both pre-training & fine-tuning frameworks and prompt tuning strategies designed for downstream adaptation:

- DGI (Veličković et al., 2018b) a self-supervised pre-training approach that learns node representations by maximizing the mutual information between local node (patch) embeddings and a high-level global summary of the graph, thereby capturing both local and global structural context in an unsupervised manner.
- GRAPHPROMPT (Liu et al., 2023) unifies pre-training and downstream tasks into a common subgraph similarity learning template. It uses a learnable prompt during pre-training to guide extraction of task-relevant knowledge and tunes the prompt during adaptation to adjust aggregation behavior so that the model better leverages pre-trained representations for specific tasks such as node and graph classification.
- HECO (Wang et al., 2021a) proposes a co-contrastive learning mechanism for heterogeneous graphs by jointly contrasting node representations from network schema and meta-path views so that the views collaboratively supervise each other to capture both local typed structure and high-order semantics.
- TGN (Rossi et al., 2020) introduces a general framework for continuous-time dynamic graphs. It maintains a memory vector for each node that is updated upon interactions to capture evolving history, and combines temporal encoding with event-driven message propagation to produce time-aware node embeddings reflecting both structural and temporal dependencies.
- DDGCL (Tian et al., 2021) proposes a debiased contrastive learning objective for dynamic graphs that alleviates sampling bias in instance discrimination by adjusting how positive and negative temporal views are contrasted.
- DSSL (Xiao et al., 2022) is a self-supervised framework for non-homophilous graphs that decouples a node's diverse neighborhood context by imitating the generative process of its neighbors with latent semantic variables, jointly capturing global cluster structure and local structural roles without relying on handcrafted augmentations.

A.4 DATASETS DESCRIPTION

In this section, we provide comprehensive descriptions of the datasets used in this benchmark, summarized in Table 1.

- *Cora*[2] is a citation network representing papers in the computer science domain, first introduced by (Chen et al., 2023). In this graph, nodes utilize titles and abstracts as raw features, edges indicate citation relationships, and labels correspond to paper categories. We employ the version

[2] https://github.com/LechengKong/OneForAll/blob/main/data/single_graph/Cora/cora.pt

provided by (Liu et al., 2024), which reformats these text features as node features and generates detailed descriptions using GPT-3.5-Turbo (ChatGPT).

- *Pubmed*[3] is a citation network representing papers in the biomedical domain, first introduced by (Chen et al., 2023). Similar to *Cora*, nodes utilize titles and abstracts as raw features, and edges indicate citation relationships. The labels correspond to three specific categories: diabetes mellitus experimental, diabetes mellitus type 1, and diabetes mellitus Type 2. We employ the version provided by (Liu et al., 2024), which uses the same processing procedure as *Cora*.
- *ogbn-arxiv*[4] is a citation network representing Computer Science arXiv papers, provided by the Open Graph Benchmark (OGB) (Hu et al., 2020a). In this graph, nodes correspond to papers and edges denote citation relationships. The task is to predict the primary subject area of each paper across 40 distinct categories (e.g., cs.AI).
- *ogbn-mag*[5] is a heterogeneous network composed of a subset of the Microsoft Academic Graph (MAG), provided by the OGB benchmark. It contains four types of entities: papers, authors, institutions, and fields of study. The edges represent relationships, including "affiliated with", "writes", "cites", and "has topics". In this graph, paper nodes are associated with 128-dimensional word2vec feature vectors. For other entity types that lack raw features, we employ MetaPath2Vec (Dong et al., 2017) to generate their input embeddings. The task is to predict the venue (conference or journal) of each paper node across 349 distinct categories.
- *ACM*[6] is a heterogeneous citation network provided by PyTorch Geometric (PyG), sourced from (Lv et al., 2021). It contains four types of entities: papers, authors, subjects, and terms. The edges represent relationships, including "cites", "writes", "belongs to", and "mentions". The task is to predict the primary category of each paper across three distinct classes: Database, Wireless Communication, and Data Mining.
- *DBLP*[7] is a heterogeneous bibliography network provided by PyG, sourced from (Lv et al., 2021). It contains four types of entities: authors, papers, terms, and venues. The edges represent relationships, including "writes", "mentions", and "published in". The task is to predict the research area of each author node across four distinct categories: Database, Data Mining, Artificial Intelligence, and Information Retrieval.
- *Reddit*[8] is a temporal interaction network provided by PyG, sourced from (Kumar et al., 2019). It contains two types of entities: users and posts. The edges represent relationships, including "interacts". The task is to predict the state of each user node, specifically whether the user is banned.
- *Wikipedia*[9] is a temporal interaction network provided by PyG, sourced from (Kumar et al., 2019). It contains two types of entities: users and pages. The edges represent relationships, including "edits". The task is to predict the state of each user node, specifically whether the user is banned.
- *Actor*[10] is a heterophilic actor network provided by (Li et al., 2025). It contains nodes representing actors extracted from Wikipedia. The edges indicate co-occurrence relationships on the same Wikipedia page. The task is to predict the category of each actor across five distinct classes, such as "American film actors" and "English actors".

[3] https://github.com/LechengKong/OneForAll/blob/main/data/single_graph/Pubmed/pubmed.pt

[4] https://snap.stanford.edu/ogb/data/misc/ogbn_arxiv

[5] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.OGB_MAG.html

[6] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.HGBDataset.html

[7] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.HGBDataset.html

[8] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.JODIEDataset.html

[9] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.JODIEDataset.html

[10] https://github.com/honey0219/LLM4HeG/blob/main/dataset/Actor.npz

- *Texas*[11], *Wisconsin*[12], *Cornell*[13] are heterophilic web page networks provided by (Li et al., 2025). It contains nodes representing web pages from the computer science department of the three universities. The edges represent hyperlinks between pages. The task is to predict the category of each web page across five distinct classes: student, project, course, staff, and faculty.
- *Chameleon*[14] is a Wikipedia page network provided by PyG, sourced from (Rozemberczki et al., 2021). It contains nodes representing web pages related to the topic of chameleons, and edges representing hyperlinks between them. The task is to predict the average daily traffic of each web page across five distinct categories.
- *IMDB*[15] is a heterogeneous movie network provided by PyG, sourced from (Lv et al., 2021). It comprises four entity types: movies, actors, directors, and keywords. The edges indicate relationships including "directed by", "acted in", and "described by". The task is to predict the genre of each movie node across five distinct categories: Action, Comedy, Drama, Romance, and Thriller.
- *Computers*[16] and *Photo*[17] are segments of the Amazon co-purchase graph provided by PyG, sourced from (Shchur et al., 2018). In these graphs, nodes represent goods, and edges indicate that two goods are frequently bought together. The task is to predict the respective product category for each good.
- *Amazon*[18] is an heterogeneous network sourced from (Cen et al., 2019), specifically focusing on the Electronics category. In this graph, nodes represent products, and edges denote distinct types of heterogeneous relationships, including co-viewing and co-purchasing links. The dataset incorporates product metadata such as price, sales-rank, brand, and category.
- *Amazon-HeTGB*[19] is a heterophilic product co-purchasing network provided by (Li et al., 2025). It contains nodes representing products such as books, music CDs, DVDs, and VHS tapes. The edges indicate relationships where products are frequently bought together. The task is to predict the average rating of each product across five distinct classes.
- *Products*[20] is a co-purchasing network employed by (Chen et al., 2024a), originally derived from the OGB benchmark. In this graph, nodes represent products sold on Amazon, and edges indicate co-purchasing relationships between them. The task is to predict the category of each product across 47 distinct classes.
- *Wiki*[21] is a large-scale knowledge graph provided by PyG, sourced from (Wang et al., 2021b). It contains nodes representing real-world entities derived from Wikipedia and Wikidata, while edges denote relational facts connecting them. The task is typically knowledge graph completion, aiming to predict missing links between entities.
- *FB15K-237*[22] is a knowledge graph dataset provided by the PyKEEN library, sourced from (Toutanova & Chen, 2015). It is a subset of the Freebase knowledge base, constructed to remove inverse relations found in the original FB15k dataset to prevent data leakage. In this graph, nodes represent general knowledge entities, and edges denote semantic relationships. The task is link prediction, aiming to infer missing relationships between entities.

[11] https://github.com/honey0219/LLM4HeG/tree/main/dataset/Texas.npz

[12] https://github.com/honey0219/LLM4HeG/blob/main/dataset/Wisconsin.npz

[13] https://github.com/honey0219/LLM4HeG/tree/main/dataset/Cornell.npz

[14] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.WikipediaNetwork.html

[15] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.HGBDataset.html

[16] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Amazon.html

[17] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Amazon.html

[18] https://github.com/SY1706203/GATNE/tree/master/dataset/amazon

[19] https://github.com/honey0219/LLM4HeG/tree/main/dataset/Amazon.npz

[20] https://utexas.app.box.com/s/i7y03rzm40xt9bjbaj0dfdgxeyjx77gb/file/1441818321356

[21] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.Wikidata5M.html

[22] https://pykeen.readthedocs.io/en/stable/api/pykeen.datasets.FB15k237.html

- *WN18RR*[23] is a knowledge graph dataset provided by the PyKEEN library, sourced from (Dettmers et al., 2018). It is a subset of WordNet, derived from the WN18 dataset by removing inverse relations. In this graph, nodes correspond to English words or synsets, and edges represent lexical relationships such as hypernyms and hyponyms. Similar to *FB15K-237*, the primary task is link prediction.
- *NELL*[24] is a knowledge graph dataset provided by PyG, sourced from (Carlson et al., 2010) and processed by (Yang et al., 2016). It is a subset of data extracted from the Never-Ending Language Learning project. In this graph, nodes represent entities, and edges indicate semantic relationships between them. The task is to predict the category of each entity node across distinct classes.
- *T-Finance*[25] is a financial transaction network sourced from (Tang et al., 2022). In this graph, nodes represent accounts, and edges denote transaction relationships between them. The node features consist of various transaction attributes and statistical information. The task is anomaly detection, aiming to classify each account node as either normal or anomalous (fraudulent).
- *Elliptic*[26] is a Bitcoin transaction network provided by PyG, sourced from (Weber et al., 2019). In this graph, nodes correspond to confirmed Bitcoin transactions, and directed edges represent the flow of Bitcoin currency between them. Each node is associated with 166 features, including time steps, local transaction information, and aggregated neighbor features. The task is to classify transactions into two categories: licit or illicit.
- *DGraph*[27] is a large-scale financial social network provided by PyG, sourced from (Huang et al., 2022). It represents a real-world graph from Finvolution, where nodes correspond to users, and edges denote 11 distinct interaction relationships such as emergency contacts. The nodes are associated with anonymized features reflecting user profiles and behavioral information. The task is anomaly detection, aiming to distinguish between normal users and fraudsters.
- *HIV*[28] and *PCBA*[29] are large-scale molecular graph datasets provided by PyG, sourced from (Wu et al., 2018). In these graphs, nodes represent atoms, and edges denote chemical bonds. The task is graph classification, aiming to predict specific biological properties, such as the ability to inhibit HIV replication or biological activity against various targets.
- *COX2*[30] and *BZR*[31] are molecular graph datasets provided by PyG, sourced from (Morris et al., 2020). Each graph represents a chemical compound, where nodes correspond to atoms and edges represent chemical bonds. The task is binary graph classification, predicting whether the molecule is active or inactive against specific receptors.
- *PROTEINS*[32] and *ENZYMES*[33] are bioinformatics datasets provided by PyG, sourced from (Morris et al., 2020). These graphs represent protein tertiary structures, where nodes correspond to secondary structure elements (helices, sheets, or turns) and edges indicate spatial proximity or

[23] https://pykeen.readthedocs.io/en/stable/api/pykeen.datasets.WN18RR.html

[24] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.NELL.html

[25] https://github.com/squareRoot3/Rethinking-Anomaly-Detection/tree/master

[26] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.EllipticBitcoinDataset.html

[27] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.DGraphFin.html

[28] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.MoleculeNet.html

[29] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.MoleculeNet.html

[30] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.TUDataset.html

[31] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.TUDataset.html

[32] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.TUDataset.html

[33] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.TUDataset.html

chemical connections. The task is graph classification, predicting the functional category of the proteins.

- *ogbn-proteins*[34] is an undirected, weighted protein-protein association network sourced from the OGB benchmark. In this graph, nodes represent proteins across eight different species, and edges denote biologically meaningful associations, such as physical interactions or co-expression. Each edge is associated with 8-dimensional features representing the confidence scores of different association types. The task is multi-label node classification, aiming to predict the presence of 112 distinct protein functions.

B Data and Experimental Protocols

B.1 Data Preparation

In this section, we elaborate on the data pre-processing protocols introduced in Section 4. To bridge the gap between diverse raw data formats (e.g., temporal streams, heterogeneous schemas) and the unified input requirements of GFMs, we implement a standardized data preparation pipeline. This pipeline converts all datasets into a unified graph object while preserving structural and semantic integrity.

Unified data object interface. Regardless of the original format, every pre-processed graph is encapsulated in a unified data object. Let N be the number of nodes and M be the number of edges. The object consists of the following standardized attributes:

- `x`: Node feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$. Defaulted to one-hot vector if unavailable.
- `edge_index`: Sparse adjacency matrix $\mathbf{E} \in \mathbb{Z}^{2 \times M}$.
- `edge_attr`: Edge feature matrix $\mathbf{E}_{attr} \in \mathbb{R}^{M \times D}$. Defaulted to zero vectors if unavailable.
- `node_type`: Integer Tensor $\mathbf{T}_{node} \in \mathbb{Z}^N$ mapping each node to its original heterogeneous type ID. Defaults to a zero-filled tensor for homogeneous graphs.
- `edge_type`: Integer Tensor $\mathbf{T}_{edge} \in \mathbb{Z}^M$ mapping each edge to its relation type ID. Defaults to a zero-filled tensor if relation types are unspecified.
- `batch`: Integer Tensor $\mathbf{B} \in \mathbb{Z}^N$ mapping nodes to graph indices within a mini-batch (only present in multi-graph datasets).
- `raw_texts`: A list of length N containing raw text for each node (e.g., titles, abstracts, or type texts). Defaulted to 'N/A' if unavailable.
- `relation_texts`: A list of length K (number of edge types) describing the semantics of relations (e.g., "cites", "co-authors"). Defaults to 'to' if unavailable.
- `label_names`: A list of natural language names of target classes.
- `label_descs`: A list of detailed textual descriptions of target classes to support zero-shot inference.
- `name`: A unique string denoting the name of the dataset.

Notably, we only extract the necessary attributes from this unified object to serve as inputs, tailored to the specific requirements of each model. A distinct exception is GraphCLIP, which ignores the original feature matrix and instead generates `x` by encoding the `raw_texts` of each node using the all-MiniLM-L6-v2 Sentence Transformer, deriving embeddings via mean pooling of the last hidden states. Table 8 provides specific examples of these `raw_texts` and `relation_texts` attributes across varying domains.

Cross-dataset feature-dimension alignment. The feature width F in the unified object is allowed to vary across datasets. Before a feature matrix is passed to a GFM, we follow the dimension-alignment strategy adopted in the corresponding original implementation and transform it to the method-specific shared input dimension reported in Table 9. Specifically, MDGPT, SAMGPT,

[34] https://ogb.stanford.edu/docs/nodeprop

Table 8: Examples of standardized `raw_texts` and `relation_texts` encapsulated in our unified data objects.

Dataset	`raw_texts`	`relation_texts`
Cora, ogbn-arxiv	Title: "Further facts..." Abstract: "Previous results..." / Title: "Spring embedders..." Abstract: "Force-directed..."	cites
ACM	author, term, subject, paper	cites, mentions, writes, belongs_to, ...
DBLP	venue, author, term, paper	published_in, mentions, writes, ...
Reddit	user, post	interacts, rev_interacts
Texas, Cornell, Wisconsin	Content: Automated Theorem Proving Group... / Content: Steve Seitz's View Interpolations...	links_to
IMDB	movie, director, actor, keyword	directed_by, stars, described_by, ...
Amazon-HeTGB	Name/title: I Ching Workbook. Salesrank: 160648. Reviews: rating: 5 votes: 1...	co-purchasing
Amazon, Photo, Computers	product	co-purchasing, also_bought, also_viewed
HIV, COX2, PROTEINS, ENZYMES	N/A	to
FB15K-237	Entity: /m/06n90	/film/film/film_art_direction_by, ...
NELL, Elliptic	entity	relation; transaction

MDGFM, GFT, and UNIGRAPH2 use PCA to project raw numerical node features to 50 dimensions; G2P2 uses PCA with an output dimension of 128; and GCOPE uses SVD to align arbitrary-dimensional inputs to 100 dimensions. GRAPHCLIP does not consume the raw numerical features: it encodes node text using the ALL-MINILM-L6-V2 Sentence Transformer to obtain fixed 384-dimensional representations. Thus, although the original feature dimensions are dataset dependent, the backbone of each GFM always receives a consistent input width across its pre-training and downstream datasets.

For memory-efficient alignment of high-dimensional sparse features, we represent the input matrix in COO format and perform the PCA/SVD projection without materializing the original dense matrix. This treatment is used for large identity or one-hot feature matrices, including *ogbn-proteins*, *FB15K-237*, *WN18RR*, and *Reddit*; the natively sparse features of *NELL*; and homogenized heterogeneous graphs such as *ACM*, *DBLP*, *IMDB*, and *Amazon*, whose zero-padded feature matrix contains only a small fraction of nonzero entries. We dynamically use sparse storage when both the number of nodes exceeds 10,000 and the raw feature dimension exceeds 1,000. This optimization changes only the storage and projection procedure; the resulting dense embeddings retain the same target dimensions and are subsequently processed identically to other datasets.

Heterogeneous graph homogenization. We flatten heterogeneous graphs into a homogeneous format to ensure compatibility with standard backbones. This process involves specific strategies for feature alignment and structural merging. Since different node types often possess varying feature dimensions or lack features entirely, we apply a rigorous alignment strategy:

- **General initialization:** For specific node types (normally one) lacking intrinsic attributes, we initialize their features using unique one-hot vectors based on node identities.
- **Specifics for ogbn-mag:** In *ogbn-mag*, we incorporate pre-trained 128-dimensional metapath2vec embeddings to create feature vectors. Specifically, "paper" node features are concatenated with these embeddings (yielding 256 dimensions), while auxiliary nodes ("author", "institution", "field of study") exclusively use the metapath2vec embeddings instead of high-dimensional sparse one-hot vectors.
- **Zero-padding alignment:** We calculate the maximum feature dimension D_{max} across all node types. Feature vectors with dimension $d < D_{max}$ are zero-padded to $\mathbb{R}^{D_{max}}$. Finally, all features are concatenated into a global matrix $\mathbf{X} \in \mathbb{R}^{N_{total} \times D_{max}}$.

Then we compute cumulative node offsets for each node type. Local edge indices are converted to global indices by adding the respective offsets. Furthermore, we consolidate multiple relations between the same pair of nodes into a single composite edge. For instance, in the *Amazon*, if Product A is linked to Product B via both "also_bought" and "also_viewed" relations, these are merged into a single edge with the composite relation text "also_bought, also_viewed" and assigned a unique `edge_type` ID.

Temporal graph snapshot construction. For dynamic graphs provided as temporal event streams, we construct a static snapshot encompassing historical information:

- **Structural aggregation:** We aggregate all interactions over the timeline. To support standard message passing, we treat the graph as undirected by generating reverse edges for every interaction, explicitly distinguishing semantics in `relation_texts` (e.g., distinguishing "interacts" from "rev_interacts" in the *Reddit*).
- **Latest-state labeling:** Since node labels may evolve, we adopt a "latest-priority" policy. For any node u, its label y_u is determined by the label associated with its last interaction timestamp t_{max}.

Graph construction policy and label management. We enforce a task-dependent and dataset-specific graph construction policy:

- **Undirected:** For most datasets, graphs are symmetrized to facilitate bi-directional information flow during message passing.
- **Directed preservation:** For datasets with inherent asymmetric semantics, including *Wiki*, *FB15K-237*, *WN18RR*, and *DGraph*, we strictly preserve edge directionality throughout the pipeline, as the meaning of $u \to v$ is fundamentally different from $v \to u$. Specifically, an adaptation is applied to *DGraph* during the node classification task: while the global graph structure is kept directed, the sampled subgraphs are symmetrized (converted to undirected) to enhance local neighborhood aggregation.

To enhance flexibility and prevent data leakage, target labels $\mathbf{y}$ are excluded from the unified data object interface. Instead, ground-truth labels and their indices are persisted in local storage. The alignment of $\mathbf{y}$ adapts to the specific task level:

- **Node-level:** Labels align with global node indices. For heterogeneous graphs where valid labels exist only on specific target types (e.g., "Paper" in *ogbn-mag*), indices corresponding to non-target nodes are filled with `-1`. This value acts as a mask token, ensuring the loss function is computed strictly on valid target nodes.
- **Edge-level:** Labels correspond to the indices of edges within the global edge list.
- **Graph-level:** Labels correspond to the graph-level `batch` ID.

This decoupling strategy allows the same unified graph structure to serve multiple diverse downstream tasks without modification.

Dataset-Specific adaptations. We further introduce multi-granularity tasks to bridge the discrepancy between the rich raw data labels and the simplified standard protocols:

- *Elliptic:* Although the standard protocol is binary (licit vs. illicit), the raw data includes a third 'others' category. We thus construct both 2-class and 3-class tasks to leverage the full label set.
- *DGraph:* While the original benchmark focuses on binary fraud detection, the raw data contains 4 distinct label indices, conceptually mapped to 3 semantic roles (Normal, Fraudster, Background Node). We extend this to include 2-class, 3-class, and 4-class tasks to evaluate multi-granularity generalization.

B.2 Dataset Sampling

In this section, we elaborate on the data sampling protocols introduced in Section 4.

Few-shot task construction. The datasets used for pre-training are *unlabeled* and restricted to those indicated in Table 1. For each downstream node, edge, or graph classification dataset, we randomly split the labeled instances into a *training pool* (20%) and a *test set* (80%). A K-shot task is constructed by sampling K labeled instances per class from the training pool, where $K = 1$ or 5. We repeat this procedure 50 times to form 50 few-shot episodes per dataset and evaluate each episode with five random seeds, resulting in $50 \times 5 = 250$ runs per dataset and shot setting. We report the mean and standard deviation of Accuracy and Macro-F1 across these runs.

Pre-training. Since different GFMs employ distinct input processing mechanisms, we implement model-specific sampling strategies:

- **Multi-domain subgraph batching:** Models (MDGPT, MDGPF, SAMGPT, G2P2, GFT, UNIGRAPH2) accept whole-graph inputs. To handle the scale discrepancy between large single-graph datasets (e.g., *ogbn-arxiv*) and multi-graph datasets (e.g., *HIV*), and to enable cross-domain attention within a single training step, we adopt a novel strategy: 1) *Decomposition:* Large single graphs are decomposed into subgraphs by `NeighborLoader` from PyG, while individual graphs in multi-graph datasets are treated directly as atomic units without further splitting. 2) *Shuffling and merging:* We maintain a mixed pool of subgraphs from all domains. During each iteration, subgraphs from diverse sources are randomly sampled, shuffled, and concatenated into a single large "super-graph" (batch). 3) *Size control:* To manage memory constraints given the variance in subgraph sizes, we control the batch size by enforcing a maximum node budget N_{max} per batch, rather than a fixed number of graphs.
- **Ego-Graph Construction:** For models that require node-centric structural contexts, we construct specific subgraphs for each target node:
  - GCOPE utilizes k-hop sampling to extract the immediate local neighborhood structure centered around each node. Notably, we exclude the original label-dependent padding for small subgraphs (< 10 nodes) to prevent data leakage, relying strictly on actual structural neighbors.
  - GRAPHCLIP employs Random Walk sampling to generate trace-based subgraphs, capturing both local structure and higher-order proximity.

Downstream. To handle the diverse scale of downstream datasets and varying model architectures, we implement an adaptive sampling strategy:

- **Full-Batch Processing (Small-Scale):** For datasets that fit entirely within GPU memory, we feed the full graph directly into the model without sampling. This ensures exact computation of representations using the complete adjacency matrix.
- **Mini-Batch Sampling (Large-Scale):** To mitigate Out-Of-Memory issues on large-scale datasets, we employ subgraph sampling techniques. Specifically, we utilize the `NeighborLoader` (for node tasks) and `LinkNeighborLoader` (for edge tasks) from PyTorch Geometric. This strategy is applied to: *ogbn-mag*, *Products*, *ogbn-proteins*, *T-Finance*, *DGraph*, and *Wiki*. By sampling local neighborhoods dynamically, we enable scalable training on these massive graphs.
- **Model-Specific Ego-Graph Batching:** Aligning with their pre-training paradigms, GCOPE and GRAPHCLIP operate on independent ego-graphs rather than a global graph view. For these models, we construct batches comprised of fixed numbers of pre-extracted ego-subgraphs.
- **Graph-Level Batching:** For graph classification tasks, we employ standard graph mini-batching. In each iteration, a fixed number of disjoint graphs are sampled and collated into a diagonal block matrix for parallel processing.

Discussion: Sampling Trade-offs. We compare node-centric ego-graphs against our dynamic neighborhood sampling. While node-centric methods offer deterministic memory control, they suffer from preprocessing bottlenecks and structural isolation. Our dynamic approach overcomes these by enabling efficient on-the-fly processing while preserving better global structural connectivity.

B.3 IMPLEMENTATION DETAILS

In this section, we elaborate on the full implementation details and hyperparameter configurations introduced in Section 4.

General settings Optimizer. For all experiments, we use the Adam optimizer (Kingma & Ba, 2015), except for GRAPHCLIP, which uses AdamW (Loshchilov & Hutter, 2019).

Environment. The environment in which we run experiments is:

- Linux version: 5.15.0-122-generic
- Operating system: Ubuntu 22.04.5 LTS (Jammy Jellyfish)
- CPU information: AMD EPYC 7763 (64-cores)
- Memory: 512 GB RAM

Table 9: General hyperparameter configurations.

Method	GNN Architecture					Pretraining			Downstream Adaptation		
	Backbone	Unified Input	Hidden Dim	Output Dim	Layers	LR	Epochs	Patience	LR	Epochs	Patience
GCOPE	FAGCN	100	128	128	2	1e-4	100	100	1e-4	100	100
MDGPT	GCN	50	256	256	3	1e-3	200	20	1e-2	400	10
MDGFM	GCN	50	256	256	3	1e-3	200	20	1e-2	400	20
SAMGPT	GCN	50	256	256	3	1e-3	200	20	1e-2	400	10
G2P2	GCN	128	128	128	2	2e-5	10	3	1e-2	50	50
GRAPHCLIP	GRAPHGPS (Rampášek et al., 2022)	384	1024	384	12	1e-5	30	30	1e-4	100	100
GFT	GRAPHSAGE (Hamilton et al., 2017)	50	128	128	3	1e-4	50	50	1e-3	200	10
UNIGRAPH2	GAT	50	768	768	3	1e-4	50	50	-	-	-

Unified Input denotes the dimension to which raw node features are projected before entering the backbone.
'-' indicates that the specific parameter is not applicable.

- GPU information: 4× NVIDIA L40 (48 GB). GRAPHCLIP pre-training is distributed across all 4 GPUs, while other experiments are conducted on a single GPU.
- CUDA Version: 12.4
- Python Version: 3.10

General Hyperparameters. Table 9 details the specific backbone architectures and hyperparameter configurations for each GFM across both pre-training and downstream phases.

Model-Specific Hyperparameters. Below, we detail the specific architectures and hyperparameter settings for each Graph Foundation Model (GFM). Unless otherwise stated, the output dimension matches the hidden dimension.

- GCOPE introduces one coordinator for each source dataset and assigns 0.2 as the default reconstruction weight.
- MDGPT adopts graph contrastive learning as the pretraining objective and employs a random edge augmentation strategy with a perturbation rate of 0.1.
- MDGFM adopts a contrastive objective between the original and the kNN-refined graphs, and reconstructs the graph structure via k-nearest neighbor (kNN) sparsification with $k = 15$ (consistent across both pretraining and downstream phases), applying a 0.5 dropout rate to the refined edge weights.
- SAMGPT follows the same configurations as MDGPT.
- G2P2 performs joint training using a 63M-parameter transformer (12 layers, 512 hidden dim, 8 heads) as text encoder based on lower-cased BPE with a vocabulary of 49,152. It sets the number of sampled neighbors η to 3 and the batch size β to 128, utilizing $M = 4$ learnable prompt tokens and a unified regularization weight $\lambda = 10$.
- GRAPHCLIP utilizes a frozen fine-tuned MiniLM text encoder and projects the 1024-dimensional graph embeddings into a 384-dimensional unified subspace, optimizing only the graph backbone and projector. Regarding the textual input generation, while the original implementation deploys the open-source QWEN2-72B-INSTRUCT model locally to summarize ego-subgraphs, we utilize the GPT-5-MINI API (OpenAI) with minimal reasoning enabled for efficient text generation.
- GFT configures the vector-quantized codebook with a commitment weight of 0.25 and applies orthonormal regularization with a weight of 1.0 over a maximum of 32 codes.
- UNIGRAPH2 employs a multimodal masked autoencoder with a 4-head GAT backbone and an 8-expert MoE module (top-2 selected). It minimizes a composite objective of feature reconstruction ($\gamma = 2.0$) and structural SPD prediction ($\lambda = 0.5$) under a 0.1 masking rate, performing inference without fine-tuning.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 ADAPTATION TO UNSEEN DATASETS

We report the 1-shot edge and graph classification results in Table 10, and the Macro-F1 of 1-shot node classification in Table 11. The 5-shot node classification results are shown in Tables 12 and 13,

Table 10: Evaluation of one-shot edge classification and graph classification on unseen datasets.

Methods	Edge classification						Graph classification			
	DGraph		Wiki		WN18RR		PCBA		BZR	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	**10.13**±3.16	6.50±0.99	9.35±3.01	3.21±0.19	15.45±11.60	10.85±4.39	55.09±31.92	32.54±15.72	52.73±16.49	44.72±10.30
GAT	9.55±3.56	6.16±1.05	2.65±0.62	1.71±0.15	10.60±14.10	7.91±3.93	56.65±28.70	33.97±13.34	52.36±17.69	43.63±10.93
GCOPE	OOT	OOT	OOT	OOT	**16.59**±7.25	**11.42**±3.29	48.99±38.52	28.19±19.08	54.31±19.35	45.89±12.33
MDGPT	9.92±2.38	6.85±0.95	10.72±3.67	2.79±0.21	13.16±4.71	10.08±2.01	55.54±21.39	34.57±9.90	53.88±13.74	47.26±9.46
MDGFM	8.37±2.17	5.86±0.78	11.04±2.95	3.49±0.33	13.72±3.11	10.32±1.54	**59.16**±40.50	32.23±20.18	44.55±25.60	31.17±12.87
SAMGPT	10.07±2.52	**6.87**±0.94	**11.45**±3.31	**3.58**±0.20	13.07±3.67	10.16±1.75	56.38±23.72	**34.63**±10.98	**56.06**±17.10	**48.07**±10.92
G2P2	-	-	-	-	12.62±3.51	9.16±1.59	-	-	53.30±13.91	46.78±8.22
GRAPHCLIP	-	-	-	-	15.44±4.60	11.20±2.04	-	-	54.47±15.52	44.67±8.77
GFT	9.28±3.25	6.50±1.12	OOT	OOT	12.66±6.67	9.96±1.64	52.55±31.20	34.25±16.59	43.67±24.29	33.56±15.81
UNIGRAPH2	9.63±3.75	5.55±1.01	10.10±0.21	3.00±0.16	10.33±3.76	6.92±1.69	50.34±9.42	33.04±4.36	51.57±16.08	43.79±9.16

Table 11: Macro-F1 of 1-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCN	43.39±11.32	4.49±0.27	28.72±19.65	20.24±7.38	21.59±4.32	12.17±1.70	26.94±21.92	23.42±7.91	37.96±9.95
GAT	39.05±10.94	3.39±0.25	31.13±12.89	18.38±7.29	21.96±4.18	12.13±1.85	33.67±20.34	23.95±8.20	31.99±6.60
SIMPLE-HGN	/	**6.63**±0.38	/	/	/	/	/	/	/
FAGCN	/	/	/	14.34±3.10	19.95±3.83	/	/	25.09±8.94	/
GCOPE	39.94±8.76	0.06±0.01	29.82±13.47	19.02±3.97	19.51±3.61	OOT	32.62±11.35	OOT	32.43±7.98
MDGPT	38.80±7.39	4.41±0.32	33.70±12.85	**24.58**±5.67	**25.98**±5.12	9.99±1.54	42.96±14.41	27.06±6.17	**47.88**±9.99
MDGFM	35.37±6.65	2.06±0.75	29.67±14.59	14.80±3.49	23.04±3.75	5.09±1.07	41.65±11.32	21.77±12.54	44.52±6.79
SAMGPT	47.82±8.75	3.77±0.23	35.88±12.86	21.59±5.80	25.05±5.33	**12.53**±1.58	43.41±10.06	**27.45**±6.83	47.83±10.66
G2P2	46.76±6.38	-	19.68±3.72	24.56±5.28	-	-	**46.46**±12.16	23.25±4.13	-
GRAPHCLIP	34.76±4.41	-	17.24±8.33	18.47±3.03	-	-	32.55±11.58	21.74±11.80	-
GFT	**53.34**±10.53	0.95±0.09	**37.42**±8.63	21.08±3.88	24.61±4.10	6.05±1.80	37.05±15.75	26.56±6.94	46.95±6.75
UNIGRAPH2	43.43±10.13	0.45±0.11	31.69±17.85	20.67±5.26	22.88±5.33	11.79±1.57	43.55±15.34	26.37±5.94	45.06±4.65

and the 5-shot edge and graph classification results are summarized in Table 14. These results exhibit the same qualitative patterns as those discussed in Sec. 5.1.

We additionally summarize downstream adaptation time (total and per-epoch) for 1-shot node classification in Tables 15 and 16, for 5-shot node classification in Tables 17 and 18, for 1-shot edge and graph classification in Table 19, and for 5-shot edge and graph classification in Table 20.

C.2 ADAPTATION TO SEEN DATASETS

We report the 1-shot edge and graph classification results in Table 21, the 5-shot node classification results in Table 22, and the 5-shot edge and graph classification results in Table 23. These results follow the same qualitative trends as those discussed in Sec. 5.1.

We further report downstream adaptation time for 1-shot and 5-shot node classification in Tables 24 and 25, and for 1-shot and 5-shot edge and graph classification in Tables 26 and 27, respectively.

C.3 TOPIC DOMAIN ADAPTATION

We provide the additional Setting III results (discussed in Sec. 5.1) for the three single-topic pretraining groups. Table 29 reports 1-shot node classification Macro-F1, Tables 30 and 31 report 5-shot node classification accuracy and Macro-F1, and Table 32 reports 5-shot edge/graph classification.

Table 12: Accuracy of 5-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCN	64.30±6.69	11.32±0.55	46.14±28.17	40.51±8.57	35.71±3.48	34.95±8.44	48.03±39.90	40.84±9.77	51.75±12.54
GAT	65.71±6.16	10.05±0.66	48.14±19.91	39.72±9.00	35.33±3.28	**43.66**±5.16	59.05±33.99	41.20±9.92	49.96±13.96
SIMPLE-HGN	/	**16.51**±1.25	/	/	/	/	/	/	/
FAGCN	/	/	/	26.61±5.65	31.50±2.78	/	/	38.56±10.35	/
GCOPE	59.13±5.40	0.25±0.06	51.77±18.38	29.71±8.06	25.41±2.19	OOT	35.86±13.81	OOT	42.44±11.57
MDGPT	53.58±6.27	11.93±0.64	54.48±18.49	**41.33**±5.26	36.33±3.26	25.61±5.47	58.90±17.82	38.17±7.42	54.00±9.61
MDGFM	51.71±7.34	OOM	49.81±19.18	23.56±5.66	31.70±3.50	8.97±3.85	54.14±12.99	39.44±13.73	50.98±5.72
SAMGPT	62.21±4.90	11.41±0.81	**58.08**±14.24	38.94±5.48	**37.86**±3.59	31.24±6.93	60.00±10.20	40.57±10.19	**54.51**±11.51
G2P2	56.55±5.41	-	43.72±8.80	37.64±5.23	-	-	67.44±11.81	**43.52**±8.90	-
GRAPHCLIP	45.58±4.54	-	43.65±12.48	27.49±4.44	-	-	45.42±25.80	36.43±14.03	-
GFT	**65.73**±4.76	3.76±0.64	56.74±10.69	35.97±5.31	33.21±2.72	20.84±3.54	67.83±13.97	42.22±7.73	52.96±5.27
UNIGRAPH2	57.99±7.02	3.61±1.94	51.36±19.89	32.91±6.61	32.93±3.33	26.63±13.32	**69.07**±13.72	37.61±7.95	50.45±3.19

Table 13: Macro-F1 of 5-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCN	63.59±7.28	10.51±0.29	31.07±13.99	32.90±5.64	34.91±3.53	17.84±3.39	31.14±23.50	29.72±4.80	41.13±11.52
GAT	65.14±6.32	9.08±0.29	33.42±9.54	30.96±6.43	34.68±3.40	**24.08**±1.39	40.97±22.11	30.11±4.85	33.54±6.99
SIMPLE-HGN	/	**12.11**±0.56	/	/	/	/	/	/	/
FAGCN	/	/	/	19.88±2.61	30.13±3.14	/	/	27.58±5.27	/
GCOPE	59.15±5.35	0.06±0.01	35.54±8.85	24.29±4.88	23.83±1.96	OOT	28.58±8.11	OOT	30.93±7.43
MDGPT	52.84±6.44	9.39±0.29	36.50±8.66	**33.60**±3.56	35.19±3.19	13.15±1.92	43.33±10.49	28.41±3.62	**52.04**±9.65
MDGFM	50.99±7.71	OOM	34.26±9.04	18.46±3.87	30.51±4.16	4.77±1.51	41.47±7.71	28.49±6.78	47.66±4.13
SAMGPT	62.16±4.87	8.91±0.64	38.56±5.89	31.67±3.65	**37.22**±3.77	15.60±1.94	44.89±5.95	29.47±4.99	51.52±10.31
G2P2	55.01±5.34	-	21.65±2.99	31.80±3.64	-	-	49.52±7.43	25.35±3.06	-
GRAPHCLIP	45.09±4.92	-	21.51±4.32	21.32±2.50	-	-	31.97±14.00	21.76±5.42	-
GFT	**65.80**±4.65	2.25±0.23	**38.85**±5.23	29.69±3.53	31.70±2.94	9.93±1.28	**50.68**±9.18	**31.29**±4.12	51.15±4.45
UNIGRAPH2	56.17±6.87	0.57±1.60	33.08±10.86	27.12±4.17	29.96±3.68	12.77±1.51	40.41±5.57	26.69±4.16	47.98±2.17

Table 14: Evaluation of 5-shot edge classification and graph classification on unseen domains.

Methods	Edge classification						Graph classification			
	DGraph		Wiki		WN18RR		PCBA		BZR	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	10.31±1.77	7.68±0.78	**21.49**±3.38	**7.40**±0.19	20.18±13.60	**20.14**±4.81	55.24±30.77	32.74±15.69	60.22±13.99	49.94±9.70
GAT	10.47±1.75	7.77±0.80	10.96±1.28	4.93±0.18	12.54±10.60	16.66±3.19	57.48±24.48	34.82±12.64	56.33±13.48	49.00±9.42
GCOPE	OOT	OOT	OOT	OOT	18.25±4.67	14.35±2.53	47.70±32.64	28.97±16.13	41.41±16.23	38.46±13.22
MDGPT	**10.85**±1.90	**8.02**±0.97	14.00±2.58	5.34±0.23	18.43±3.62	16.03±2.66	**58.31**±16.94	**36.25**±7.37	53.97±11.76	50.35±9.68
MDGFM	9.92±2.17	6.97±0.84	11.32±2.96	3.47±0.40	17.32±2.86	13.79±1.60	55.27±27.36	33.52±13.25	37.26±19.41	30.48±11.90
SAMGPT	10.43±2.07	7.62±0.84	15.24±2.48	5.92±0.20	18.14±3.47	14.85±1.75	56.40±14.87	35.67±6.41	**65.39**±10.28	**57.64**±7.05
G2P2	-	-	-	-	16.90±2.89	13.50±1.50	-	-	60.89±9.65	54.87±6.82
GRAPHCLIP	-	-	-	-	**20.20**±4.30	15.19±1.78	-	-	53.72±17.44	44.88±10.52
GFT	10.21±2.00	7.17±0.88	OOT	OOT	16.29±5.48	12.78±1.70	52.13±15.91	33.70±9.22	62.70±14.25	52.18±10.83
UNIGRAPH2	10.22±6.11	7.64±0.89	11.13±0.24	5.02±0.20	15.35±8.76	11.93±1.26	51.66±13.08	33.38±6.00	62.46±10.80	49.78±6.20

Table 15: Time of downstream tuning for 1-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCOPE	4982.84	145726.60	3609.70	2399.45	2040.97	OOT	10597.74	OOT	30164.69
MDGPT	996.96	21295.45	578.18	559.66	621.26	41059.80	1677.46	6052.41	2246.63
MDGFM	4276.28	68365.23	520.57	571.15	977.57	206484.25	5429.48	4385.13	9203.33
SAMGPT	3100.42	37458.71	1302.22	1395.42	1413.17	101031.49	4254.17	7890.00	2561.89
G2P2	268.5	-	284.01	284.77	-	-	3150.19	1928.43	-
GRAPHCLIP	1143.31	-	549.77	603.42	-	-	1578.20	10320.75	-
GFT	1264.49	126155.83	5586.78	394.60	363.13	157924.23	41145.66	23168.21	63876.06
UNIGRAPH2	1778.73	107988.18	683.08	214.92	593.02	138796.78	22396.36	19710.11	104354.87

Table 16: Time/epoch of downstream tuning for 1-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCOPE	0.9966	29.1453	0.7219	0.4799	0.4082	OOT	2.1195	OOT	6.0329
MDGPT	0.0499	2.3057	0.0289	0.0280	0.0311	5.4761	1.1796	0.3600	2.2443
MDGFM	0.2952	4.5010	0.1541	0.0990	0.0744	12.7894	1.3986	0.6626	2.5256
SAMGPT	0.1554	2.7046	0.0721	0.0707	0.0711	6.5106	1.2225	0.4005	2.4340
G2P2	0.1074	-	0.1136	0.1139	-	-	1.2601	0.7714	-
GRAPHCLIP	0.2287	-	0.1100	0.1207	-	-	0.3156	2.0642	-
GFT	0.0903	12.6156	0.4422	0.0244	0.0235	16.6922	8.5135	2.8209	14.2089
UNIGRAPH2	-	-	-	-	-	-	-	-	-

Table 17: Time of downstream tuning for 5-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCOPE	4967.40	145716.29	3583.13	2419.86	2036.05	OOT	10599.70	OOT	30412.11
MDGPT	91.60	30592.08	573.91	559.37	620.14	29485.19	1983.93	6030.87	2309.75
MDGFM	4103.18	OOM	549.41	582.99	998.64	195360.13	4520.15	5207.33	11524.04
SAMGPT	3107.38	97086.74	1462.31	1375.35	1276.83	105312.52	4113.18	7786.37	2616.61
G2P2	279.28	-	284.19	278.05	-	-	3365.97	1920.82	-
GRAPHCLIP	1234.62	-	597.31	675.31	-	-	1624.89	10517.09	-
GFT	68.65	153450.58	8856.30	1090.68	501.34	158267.54	33193.50	18052.72	72111.36
UNIGRAPH2	1673.26	108452.68	746.61	282.63	642.26	141652.54	21014.23	19560.55	108606.81

Table 18: Time/epoch of downstream tuning for 5-shot node classification on unseen domains.

Methods	Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
GCOPE	0.9935	29.1433	0.7166	0.4840	0.4072	OOT	2.1199	OOT	6.0824
MDGPT	0.0503	4.0784	0.0287	0.0280	0.0311	6.6558	1.2596	0.3580	2.3765
MDGFM	0.2977	OOM	0.1592	0.0988	0.0819	16.6476	1.4073	0.5761	3.0493
SAMGPT	0.1554	5.5339	0.0731	0.0688	0.0638	9.0304	1.3162	0.3964	2.6281
G2P2	0.1117	-	0.1137	0.1112	-	-	1.3464	0.7683	-
GRAPHCLIP	0.2469	-	0.1195	0.1351	-	-	0.3250	2.1034	-
GFT	0.0900	15.3451	0.5737	0.0557	0.0269	16.7534	6.8271	3.6105	14.3866
UNIGRAPH2	-	-	-	-	-	-	-	-	-

Table 19: Time and time/epoch of downstream tuning for 1-shot edge classification and graph classification on unseen domains.

Methods	Edge classification						Graph classification			
	DGraph		Wiki		WN18RR		PCBA		BZR	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCOPE	OOT	OOT	OOT	OOT	5366.42	1.0733	7849.48	1.5689	1999.33	0.3999
MDGPT	6691.48	0.6109	27480.67	2.1883	1482.20	0.0741	1171.83	0.0777	373.24	0.0254
MDGFM	16287.70	1.1188	79650.24	4.0031	12818.37	0.7522	9000.25	1.8664	194.41	0.0515
SAMGPT	14194.53	0.8092	52318.30	2.6658	5529.34	0.2765	2046.01	0.1056	413.24	0.0242
G2P2	-	-	-	-	311.16	0.1245	-	-	260.62	0.1042
GRAPHCLIP	-	-	-	-	2435.14	0.4870	-	-	1500.62	0.3001
GFT	14844.22	0.9222	OOT	OOT	342.12	0.0779	57977.69	12.4772	237.40	0.1020
UNIGRAPH2	5701.38	-	73366.01	-	426.74	-	13861.10	-	227.95	-

Table 20: Time and time/epoch of downstream tuning for 5-shot edge classification and graph classification on unseen domains.

Methods	Edge classification						Graph classification			
	DGraph		Wiki		WN18RR		PCBA		BZR	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCOPE	OOT	OOT	OOT	OOT	5377.28	1.0755	7869.82	1.5740	1990.91	0.3982
MDGPT	3239.32	0.5356	33128.83	3.1503	1467.97	0.0744	1159.48	0.0925	421.82	0.0232
MDGFM	14083.09	1.1941	78096.08	3.9079	12869.09	0.7497	9296.78	1.0705	233.65	0.0366
SAMGPT	17314.93	0.9869	53692.22	2.8210	5531.02	0.2766	1943.27	0.1037	297.63	0.0152
G2P2	-	-	-	-	310.80	0.1243	-	-	261.65	0.1047
GRAPHCLIP	-	-	-	-	2421.57	0.4843	-	-	1522.43	0.3045
GFT	14859.82	0.9757	OOT	OOT	604.05	0.1468	60763.81	12.9928	185.39	0.0813
UNIGRAPH2	5288.73	-	74826.76	-	423.24	-	14087.09	-	165.73	-

Table 21: Evaluation of one-shot edge classification and graph classification on seen datasets.

Methods	Edge classification		Graph classification					
	FB15K-237		HIV		COX2		PROTEINS	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	14.63±1.4	13.35±0.8	45.79±27.1	31.22±13.8	48.09±14.2	42.15±9.5	**58.67**±9.2	52.33±8.6
GAT	11.05±1.2	10.90±0.8	46.00±27.4	31.50±14.4	50.71±14.9	43.67±9.1	57.79±9.4	51.44±8.7
DGI	4.84±1.4	1.61±0.3	48.47±13.2	34.93±6.4	46.99±12.8	41.99±8.3	54.28±8.7	51.56±8.1
GRAPHPROMPT	5.52±1.6	1.42±0.3	50.50±9.3	**36.15**±4.2	49.30±10.9	44.18±7.1	56.83±7.0	54.32±6.3
GCOPE	1.51±0.4	1.01±0.1	**54.20**±34.7	33.51±16.2	39.04±13.6	35.05±9.3	53.27±9.5	50.48±9.9
MDGPT	15.56±1.2	12.52±0.7	47.51±15.2	34.26±8.3	53.06±12.2	45.75±6.7	55.51±9.9	48.87±8.8
MDGFM	12.26±1.2	10.28±1.0	42.45±29.9	28.78±15.4	54.47±22.8	38.67±11.7	56.00±9.0	49.44±9.9
SAMGPT	14.91±1.5	12.05±1.1	46.21±17.8	33.43±9.6	**55.53**±12.0	**47.63**±7.1	58.02±8.6	52.25±7.7
G2P2	**17.26**±1.4	**14.23**±0.8	49.36±15.3	35.30±7.7	52.35±13.1	45.84±8.8	53.48±9.3	49.26±9.6
GRAPHCLIP	2.68±0.5	1.31±0.2	47.77±20.4	33.69±9.9	49.16±13.3	42.66±7.6	57.00±7.6	**55.26**±7.6
GFT	2.00±0.4	1.62±0.2	46.61±34.7	29.30±18.6	49.01±21.4	37.45±13.1	51.52±10.1	44.56±9.5
UNIGRAPH2	1.41±0.7	1.02±0.0	48.56±13.7	32.11±6.5	49.14±15.8	41.65±9.1	52.30±10.7	48.56±9.3

Table 22: Evaluation of 5-shot node classification on seen domains.

Methods	Cora		ACM		Reddit		Wisconsin		Photo		Elliptic	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	69.86±4.93	68.39±4.21	52.49±12.80	44.46±16.02	51.02±20.66	35.75±10.64	42.85±4.74	31.75±3.62	76.09±7.03	74.45±6.73	51.90±12.45	42.21±7.92
GAT	68.57±4.38	67.55±4.03	39.28±7.90	26.88±11.74	49.74±15.65	35.81±7.48	40.35±4.80	30.58±4.14	76.11±4.54	75.25±3.57	50.43±14.12	40.17±8.47
SIMPLE-HGN	/	/	66.31±20.71	62.29±26.23	/	/	/	/	/	/	/	/
FAGCN	/	/	/	/	/	/	**59.42**±9.09	47.42±6.96	/	/	/	/
DGI	58.70±6.05	55.59±5.33	42.48±10.45	38.31±11.38	50.85±9.54	36.41±4.31	31.93±7.41	25.35±5.89	56.69±7.18	55.48±6.59	44.95±14.57	35.26±10.62
GRAPHPROMPT	**72.78**±2.65	**69.58**±2.71	67.97±12.91	66.78±14.02	49.66±18.04	35.10±8.83	38.38±4.22	31.08±4.00	66.44±6.83	66.02±5.80	53.45±18.15	42.73±13.81
HECO	/	/	**86.81**±1.83	**86.90**±1.84	/	/	/	/	/	/	/	/
TGN	/	/	/	/	52.43±12.74	37.22±5.87	/	/	/	/	/	/
DDGCL	/	/	/	/	53.30±11.44	37.31±5.03	/	/	/	/	/	/
DSSL	/	/	/	/	/	/	49.24±5.44	41.63±4.45	/	/	/	/
GCOPE	62.88±3.86	60.78±3.60	34.34±1.42	31.35±2.47	22.66±23.82	17.31±13.55	24.95±6.06	22.92±5.09	46.91±3.89	42.02±4.16	42.43±8.01	34.78±4.34
MDGPT	61.79±4.36	60.96±4.07	77.11±7.95	76.54±8.59	43.53±15.68	32.45±8.55	36.15±5.54	31.39±4.22	80.14±4.47	**79.30**±3.83	51.60±10.34	40.87±6.87
MDGFM	51.03±7.39	50.29±7.08	67.83±11.65	67.41±11.93	53.42±19.37	36.68±9.44	30.53±7.64	24.98±5.94	73.86±5.18	72.37±4.47	**55.92**±9.88	**44.37**±6.93
SAMGPT	62.19±5.32	61.49±4.91	72.19±6.33	71.61±6.71	49.56±14.88	35.64±7.26	45.35±7.50	38.25±5.24	**80.36**±4.63	78.58±3.99	53.12±11.07	42.66±7.21
G2P2	57.25±4.24	56.82±3.96	70.77±5.48	69.34±7.58	47.31±9.66	23.55±3.43	42.63±6.20	36.10±4.07	-	-	44.75±5.92	30.25±3.04
GRAPHCLIP	49.41±4.44	47.36±3.80	31.54±2.29	22.82±1.54	39.44±10.60	20.78±3.72	37.07±5.63	32.69±4.39	-	-	40.13±11.00	26.00±4.29
GFT	64.15±3.23	63.56±3.15	56.66±3.99	55.17±4.13	**54.38**±7.99	**38.63**±3.85	57.19±6.03	**50.36**±4.74	75.53±5.10	74.25±4.55	47.94±8.13	39.62±4.89
UNIGRAPH2	64.44±4.09	61.57±3.78	80.15±11.06	78.50±12.37	47.57±14.68	34.44±7.29	26.35±6.77	21.13±5.19	79.26±5.61	78.34±4.99	48.84±8.35	38.44±5.10

Table 23: Evaluation of 5-shot edge classification and graph classification on seen datasets.

Methods	Edge classification FB15K-237		Graph classification HIV		Graph classification COX2		Graph classification PROTEINS	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCN	**26.85**±1.55	**25.90**±0.97	46.39±28.97	30.91±15.60	58.73±11.32	50.45±7.36	57.79±8.91	55.05±8.51
GAT	24.62±1.33	24.57±0.81	48.22±20.26	34.07±10.62	57.52±10.63	50.90±7.38	58.40±8.52	56.20±8.28
DGI	8.54±1.23	9.40±0.58	57.22±17.97	38.35±7.93	51.93±10.63	45.77±6.65	56.77±8.38	52.55±7.95
GRAPHPROMPT	12.47±1.16	10.42±0.55	51.07±9.62	36.41±4.35	53.11±7.82	47.14±4.53	54.35±8.31	52.96±7.86
GCOPE	1.92±0.33	1.20±0.14	**66.47**±26.43	**41.41**±12.13	38.11±17.34	32.81±11.76	**64.15**±8.17	**63.44**±8.09
MDGPT	22.48±1.38	18.55±0.62	48.35±10.12	35.34±5.11	51.76±7.61	47.43±5.15	53.50±9.94	48.71±11.18
MDGFM	17.62±1.26	15.58±1.06	49.36±20.94	34.45±10.43	58.97±22.17	41.10±11.54	53.86±8.70	51.89±8.81
SAMGPT	22.30±1.22	18.14±0.81	51.96±13.65	36.85±6.47	**60.12**±7.84	**53.42**±6.01	57.43±8.06	55.59±7.89
G2P2	24.87±1.19	22.30±0.65	50.76±11.37	36.52±5.41	58.48±7.52	51.79±5.50	56.83±7.22	54.41±6.73
GRAPHCLIP	3.09±0.44	1.30±0.17	52.30±17.30	36.66±9.08	48.34±19.32	39.03±11.49	60.49±7.01	59.65±6.80
GFT	7.71±0.61	7.05±0.50	52.20±12.65	36.84±5.90	59.34±16.56	47.67±11.90	56.71±8.81	52.38±10.55
UNIGRAPH2	2.51±1.97	1.52±0.13	50.57±7.64	33.40±3.69	50.01±14.81	42.79±8.64	53.10±12.27	51.18±11.26

Table 24: Time and time/epoch of downstream tuning for 1-shot node classification on seen datasets.

Methods	Cora		ACM		Reddit		Wisconsin		Photo		Elliptic	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCN	31.31	0.0140	122.84	0.0636	127.02	0.1119	129.34	0.0255	828.39	0.0698	131.64	0.1707
GAT	50.53	0.0298	184.55	0.0955	159.57	0.1886	104.03	0.0406	410.57	0.0788	174.17	0.2399
SIMPLE-HGN	-	-	3764.31	0.2510	-	-	-	-	-	-	-	-
TGN	-	-	-	-	1660.39	0.1660	-	-	-	-	-	-
FAGFN	-	-	-	-	-	-	217.31	0.0232	-	-	-	-
DGI	33.32	0.0577	565.19	0.9661	2269.40	4.0743	97.84	0.0142	625.27	0.7018	306.42	0.5211
HECO	-	-	811.97	0.3248	-	-	-	-	-	-	-	-
DDGCL	-	-	-	-	5242.64	0.5243	-	-	-	-	-	-
DSSL	-	-	-	-	-	-	25.98	0.0026	-	-	-	-
GRAPHPROMPT	26.02	0.0109	108.08	0.1965	2016.28	3.6660	126.91	0.0489	763.13	1.1875	287.58	0.2498
GCOPE	1889.88	0.3780	2011.55	0.4023	4133.96	0.8268	1385.02	0.2770	2791.60	0.5583	37958.68	7.2870
MDGPT	572.78	0.0286	1598.63	0.0886	3652.43	0.1826	560.69	0.0280	1057.15	0.0553	6157.62	0.3573
MDGFM	1244.31	0.0858	2104.96	0.2524	580.64	0.1847	445.44	0.0751	799.05	0.0625	55421.48	3.7830
SAMGPT	1244.34	0.0624	3563.67	0.2179	7999.01	0.4203	1428.11	0.0729	2796.97	0.1424	25579.69	1.2908
G2P2	315.28	0.1261	415.70	0.1663	421.09	0.1684	408.96	0.1636	-	-	436.85	0.1747
GRAPHCLIP	818.55	0.1637	1008.32	0.2017	781.38	0.1563	797.49	0.1595	-	-	1423.74	0.2847
GFT	257.49	0.0138	1506.18	0.3203	4037.43	0.8261	338.40	0.1055	1228.08	0.0756	1793.17	0.3915
UNIGRAPH2	73.43	-	1877.38	-	867.44	-	107.03	-	1233.20	-	4653.17	-

Table 25: Time and time/epoch of downstream tuning for 5-shot node classification on seen datasets.

Methods	Cora		ACM		Reddit		Wisconsin		Photo		Elliptic	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCN	257.61	0.0297	246.73	0.0681	318.01	0.1295	560.74	0.0310	196.12	0.0269	171.76	0.1286
GAT	296.70	0.0464	228.42	0.1224	220.35	0.1805	195.15	0.0213	500.67	0.0357	245.24	0.1822
SIMPLE-HGN	-	-	3765.52	0.2510	-	-	-	-	-	-	-	-
TGN	-	-	-	-	1664.02	0.1664	-	-	-	-	-	-
FAGFN	-	-	-	-	-	-	282.46	0.0303	-	-	-	-
DGI	46.37	0.0594	676.29	1.0926	2811.11	4.9842	130.31	0.0253	719.83	1.0883	331.31	0.5895
HECO	-	-	812.28	0.3249	-	-	-	-	-	-	-	-
DDGCL	-	-	-	-	5264.01	0.5264	-	-	-	-	-	-
DSSL	-	-	-	-	-	-	33.75	0.0034	-	-	-	-
GRAPHPROMPT	28.18	0.0127	183.35	0.1991	2599.13	4.2819	130.49	0.0511	790.62	1.2425	327.36	0.2775
GCOPE	1884.36	0.3769	1979.95	0.3960	4175.79	0.8352	1379.19	0.2758	2788.78	0.5578	37996.33	7.3405
MDGPT	567.61	0.0288	1604.10	0.0884	3654.25	0.1827	559.22	0.0280	917.99	0.0536	5731.58	0.3593
MDGFM	1286.06	0.0867	1999.56	0.2566	575.39	0.1860	336.84	0.0581	764.06	0.0667	38802.31	3.8225
SAMGPT	1261.57	0.0631	4639.72	0.2320	7674.79	0.4112	1206.69	0.0603	3308.01	0.1654	25805.74	1.2676
G2P2	328.73	0.1315	415.54	0.1662	426.76	0.1707	403.93	0.1616	-	-	430.23	0.1721
GRAPHCLIP	892.39	0.1785	1047.32	0.2095	857.76	0.1716	952.70	0.1905	-	-	1389.87	0.2780
GFT	241.75	0.0201	1591.40	0.3318	3819.50	0.8196	385.16	0.1241	1267.71	0.0771	1441.96	0.3497
UNIGRAPH2	73.02	-	2168.43	-	861.27	-	125.94	-	1907.82	-	4226.65	-

Table 26: Time and time/epoch of downstream tuning for 1-shot edge classification and graph classification on seen datasets.

Methods	Edge classification FB15K-237		Graph classification HIV		Graph classification COX2		Graph classification PROTEINS	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCN	1220.51	0.0639	166.80	0.0414	47.53	0.0250	44.65	0.0346
GAT	923.32	0.0782	145.08	0.0715	42.43	0.0280	39.23	0.0346
DGI	4468.84	3.4212	3769.68	2.8540	43.64	0.0248	119.66	0.2176
GRAPHPROMPT	2239.06	0.2528	2495.85	2.5379	35.17	0.0640	181.10	0.3293
GCOPE	2361.41	0.4723	2029.85	0.4060	1465.00	0.2930	1518.61	0.3037
MDGPT	3073.88	0.1537	369.77	0.0267	376.72	0.0236	53.92	0.0109
MDGFM	4454.41	0.2445	1443.36	0.1282	321.10	0.0518	483.72	0.0525
SAMGPT	5435.56	0.2718	1229.38	0.0630	801.66	0.0448	1085.71	0.0557
G2P2	2013.12	0.8052	331.19	0.1325	307.83	0.1231	309.76	0.1239
GRAPHCLIP	5032.46	1.0065	2913.34	0.5827	718.89	0.1438	738.50	0.1477
GFT	6007.92	1.2729	3389.08	0.9007	410.75	0.0785	700.49	0.1712
UNIGRAPH2	5643.13	-	2835.38	-	686.40	-	1159.89	-

Table 27: Time and time/epoch of downstream tuning for 5-shot edge classification and graph classification on seen datasets.

Methods	Edge classification FB15K-237		Graph classification HIV		Graph classification COX2		Graph classification PROTEINS	
	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.	Time	Avg. T.
GCN	1002.11	0.0796	332.07	0.0303	375.00	0.0277	127.64	0.0163
GAT	779.50	0.0876	290.26	0.0269	95.51	0.0110	119.32	0.0121
DGI	4640.11	3.6991	3807.46	2.9227	56.63	0.0302	131.69	0.2394
GRAPHPROMPT	2599.97	0.2667	2651.73	2.8213	44.45	0.0808	181.88	0.3307
GCOPE	2371.46	0.4743	2014.62	0.4029	1450.19	0.2900	1508.49	0.3017
MDGPT	3394.25	0.1697	314.36	0.0335	332.59	0.0267	91.82	0.0329
MDGFM	4498.39	0.2447	1397.38	0.1435	460.08	0.0597	411.02	0.0518
SAMGPT	5450.74	0.2725	1408.20	0.0704	1156.79	0.0641	695.99	0.0393
G2P2	2013.91	0.8056	340.58	0.1362	316.66	0.1267	295.84	0.1183
GRAPHCLIP	11790.04	2.3580	3306.35	0.6613	748.74	0.1497	999.50	0.1999
GFT	10723.83	2.6556	3459.66	1.0651	469.25	0.1450	407.72	0.1432
UNIGRAPH2	5686.28	-	3548.80	-	620.54	-	1241.42	-

Table 28: Evaluation of one-shot edge classification and graph classification on unseen datasets when separately pre-trained on graphs from three different topic domains.

Pre-training topic	Methods	Edge Classification						Graph Classification			
		Finance		**Common sense**				**Molecule**			
		DGraph		Wiki		WN18RR		PCBA		BZR	
		Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
Citation (Cora, ACM, ogbn_arxiv, DBLP)	GCOPE	OOT	OOT	OOT	OOT	**17.40**±6.01	**12.82**±3.50	**62.00**±33.78	**35.19**±16.07	56.05±19.87	46.31±13.01
	MDGPT	9.66±2.65	6.34±0.95	7.69±2.25	1.73±0.22	15.18±5.18	11.29±2.57	54.51±25.18	33.57±11.75	54.50±14.94	48.42±11.09
	MDGFM	9.15±2.03	6.19±0.71	4.57±2.86	0.60±0.11	15.17±5.18	10.15±2.32	57.68±41.62	31.41±20.54	55.00±24.60	36.10±11.21
	SAMGPT	9.92±2.64	**6.79**±1.05	10.69±2.85	3.12±0.19	12.93±4.39	9.41±1.96	55.71±23.29	34.36±10.82	**57.51**±15.30	**49.86**±10.75
	G2P2	-	-	-	-	11.42±4.41	8.36±1.86	-	-	52.55±16.75	45.05±10.80
	GRAPHCLIP	-	-	-	-	14.30±2.16	11.91±1.76	-	-	51.24±20.24	41.04±11.17
	GFT	**10.43**±3.53	6.70±1.32	OOT	OOT	10.59±3.33	6.38±1.33	51.30±32.80	30.64±17.22	49.11±26.15	33.69±14.40
	UNIGRAPH2	9.45±1.89	6.57±0.79	**12.99**±3.07	**3.99**±0.73	9.89±2.95	6.18±1.24	48.53±22.92	31.54±13.49	54.02±9.05	46.86±5.83
Social (Reddit, Texas, Wisconsin, Cornell, IMDB)	GCOPE	OOT	OOT	OOT	OOT	**16.28**±5.50	**12.27**±3.35	52.98±25.36	32.88±11.90	57.18±17.15	47.36±9.60
	MDGPT	9.55±2.80	6.50±0.99	9.54±2.44	2.78±0.23	12.13±4.33	8.75±1.97	54.31±22.99	**33.89**±10.44	52.40±14.47	46.29±10.12
	MDGFM	9.90±3.36	6.02±1.10	3.78±2.65	0.44±0.10	14.28±5.65	9.39±2.71	**57.68**±41.26	31.47±20.27	49.08±22.32	35.30±10.05
	SAMGPT	10.04±2.95	**6.84**±1.05	11.09±3.04	3.53±0.19	12.86±4.44	9.17±1.93	54.31±22.66	33.88±10.40	**58.00**±14.31	**50.18**±10.03
	G2P2	-	-	-	-	13.96±6.82	8.83±2.01	-	-	52.15±14.87	44.94±8.89
	GRAPHCLIP	-	-	-	-	14.99±4.15	10.45±1.87	-	-	52.50±15.21	43.38±8.64
	GFT	**10.12**±2.83	6.82±1.10	OOT	OOT	11.30±3.83	7.79±1.65	51.32±34.97	30.69±20.97	52.87±27.41	34.53±14.20
	UNIGRAPH2	9.55±2.30	6.57±0.95	**11.85**±3.99	**3.61**±0.55	10.52±3.98	7.14±1.65	49.94±23.23	31.58±14.25	52.54±9.69	45.81±7.26
Molecule (HIV, COX2)	GCOPE	OOT	OOT	OOT	OOT	**18.49**±6.84	**12.86**±3.83	49.07±39.45	27.87±19.81	49.68±22.61	42.79±17.05
	MDGPT	9.58±2.56	6.72±0.86	9.69±2.72	2.62±0.23	12.74±4.35	9.25±2.09	53.79±22.03	**33.80**±9.79	52.26±15.28	46.26±11.58
	MDGFM	8.77±2.55	5.87±0.92	4.07±1.99	0.58±0.14	13.64±4.40	9.30±1.77	**57.66**±38.82	32.12±18.89	51.87±24.20	35.81±11.17
	SAMGPT	9.52±2.89	6.59±1.16	**10.32**±2.93	**2.84**±0.46	13.94±4.43	9.65±2.12	54.21±24.32	33.57±11.37	**61.61**±15.67	**51.85**±11.21
	G2P2	-	-	-	-	12.71±6.15	8.72±2.71	-	-	51.69±17.17	43.85±10.16
	GRAPHCLIP	-	-	-	-	12.34±3.33	8.60±1.36	-	-	49.86±17.32	42.57±10.54
	GFT	**10.04**±2.85	6.74±1.05	OOT	OOT	11.53±3.91	7.46±1.42	50.80±35.95	30.17±17.94	52.04±25.05	33.39±14.28
	UNIGRAPH2	9.66±2.24	**6.79**±0.82	10.07±2.77	2.68±0.54	10.19±3.78	6.17±1.43	49.78±23.42	31.89±13.34	53.03±14.73	44.61±7.86

Table 29: Complete Macro-F1 of 1-shot node classification on topic domain adaptation under three single-topic pre-training groups.

Pre-train	Methods	Citation		Social & Web			E-commerce	Finance		Proteins
		Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
Citation	GCOPE	32.31±6.92	0.13±0.01	34.12±14.34	15.53±4.56	20.56±2.40	OOT	25.24±15.32	OOT	35.62±8.01
	MDGPT	26.97±7.29	2.50±0.27	30.41±18.29	22.37±5.29	24.51±4.62	7.93±1.68	44.31±15.86	22.61±12.43	47.75±9.97
	MDGFM	34.22±5.32	0.68±0.46	31.32±16.20	14.41±3.80	21.30±4.50	3.74±1.29	36.99±13.74	19.96±13.08	42.25±7.24
	SAMGPT	43.93±6.98	**4.67**±0.29	31.70±14.50	18.58±3.84	25.31±5.09	**10.68**±1.48	39.67±12.38	24.16±9.78	48.17±10.18
	G2P2	45.64±7.80	-	19.57±2.75	22.52±7.01	-	-	44.46±10.41	22.49±4.98	-
	GRAPHCLIP	**47.42**±6.21	-	22.34±7.37	**28.84**±2.75	-	-	37.27±11.26	20.77±11.74	-
	GFT	41.30±8.16	1.39±0.09	**40.27**±12.14	20.61±4.03	25.17±4.23	5.97±0.83	40.13±8.63	**26.44**±6.70	47.93±5.22
	UNIGRAPH2	45.08±11.60	1.42±0.10	30.88±13.50	18.98±4.77	**25.37**±4.60	10.45±1.29	**56.89**±22.20	24.89±7.30	**48.74**±1.23
Social	GCOPE	34.87±6.58	0.12±0.01	35.95±12.68	16.49±4.66	19.08±2.74	OOT	31.05±12.66	OOT	37.93±8.80
	MDGPT	42.26±8.14	**4.98**±0.34	34.23±11.88	23.05±5.45	**27.23**±4.78	10.48±1.48	39.57±16.44	26.05±5.82	**49.14**±10.66
	MDGFM	33.21±5.33	0.46±0.29	32.00±17.13	13.24±3.34	22.45±3.60	2.30±0.69	41.16±16.09	22.35±12.22	43.88±6.96
	SAMGPT	43.25±9.52	4.87±0.28	32.41±14.12	19.94±4.35	27.22±4.69	**12.00**±1.44	40.20±9.35	**26.59**±6.43	47.57±10.82
	G2P2	**47.68**±8.89	-	24.33±7.10	**23.30**±7.01	-	-	44.93±10.65	23.86±5.51	-
	GRAPHCLIP	33.34±4.13	-	20.90±8.22	15.95±3.03	-	-	36.37±11.61	22.46±10.94	-
	GFT	41.95±8.06	1.43±0.12	**37.76**±14.53	22.58±3.15	25.58±4.17	5.90±0.74	45.42±13.85	26.12±7.16	47.97±5.18
	UNIGRAPH2	45.26±10.87	1.11±0.09	29.73±14.36	19.83±4.66	26.05±4.39	9.82±1.24	**69.78**±27.96	24.98±7.83	48.21±2.28
Molecular	GCOPE	29.07±7.15	0.04±0.01	20.93±19.29	12.70±4.29	18.15±1.96	OOT	33.68±9.56	OOT	32.90±7.45
	MDGPT	45.09±8.74	**4.86**±0.39	34.19±7.92	**24.29**±5.30	26.42±5.18	12.01±1.50	42.37±20.21	25.57±5.99	48.35±10.32
	MDGFM	35.94±7.62	0.47±0.15	32.79±15.61	13.63±3.38	22.94±4.49	4.72±1.55	37.82±13.98	24.42±8.82	42.64±7.34
	SAMGPT	**47.82**±8.04	4.56±0.35	33.34±13.74	18.86±4.22	**26.76**±4.53	**12.33**±1.55	41.36±10.93	**27.49**±8.53	**50.07**±9.96
	G2P2	43.68±7.47	-	25.65±7.89	22.70±6.91	-	-	45.55±10.43	27.36±6.02	-
	GRAPHCLIP	30.30±3.20	-	17.07±8.59	15.04±3.00	-	-	35.04±11.69	21.05±12.28	-
	GFT	42.70±8.84	1.37±0.12	**38.85**±14.01	22.03±3.99	25.21±3.84	5.93±0.69	49.99±17.65	26.70±7.45	47.75±5.34
	UNIGRAPH2	43.13±10.62	0.70±0.06	30.06±13.03	17.86±3.61	25.19±3.68	6.54±0.81	**61.67**±23.06	25.99±7.30	48.28±3.33

Table 30: Complete accuracy of 5-shot node classification on topic domain adaptation under three single-topic pre-training groups.

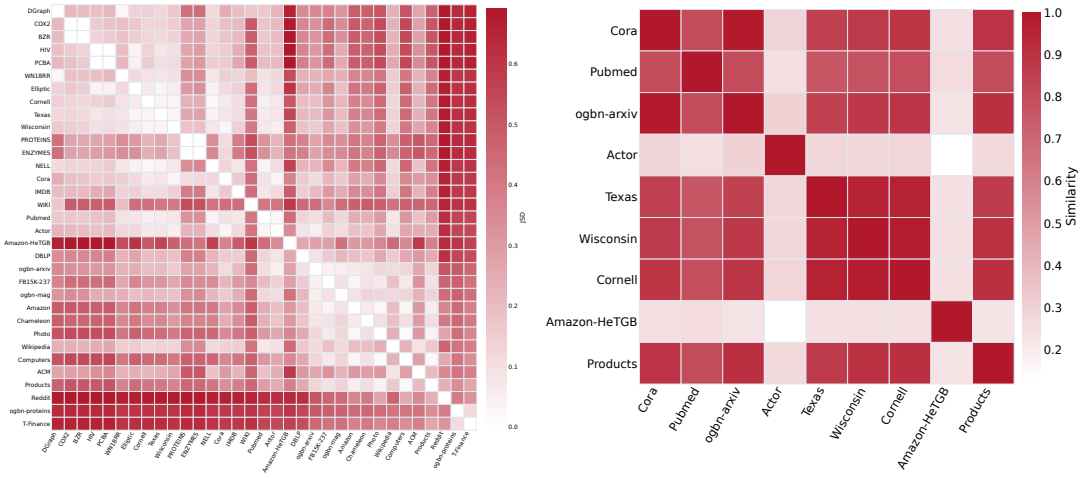
Pre-train	Methods	Citation		Social & Web			E-commerce	Finance		Proteins
		Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
Citation	GCOPE	54.25±5.49	0.27±0.04	52.11±18.57	29.04±5.83	25.18±1.82	OOT	33.29±18.15	OOT	47.01±12.49
	MDGPT	54.42±8.44	7.97±0.93	54.09±23.14	38.60±6.31	34.02±3.00	17.34±5.20	59.92±19.35	37.78±11.71	**54.18**±10.72
	MDGFM	43.25±6.96	OOM	40.03±26.83	24.88±7.61	27.31±3.64	5.54±1.50	55.98±15.99	38.77±13.40	52.01±8.78
	SAMGPT	60.22±5.65	**11.56**±0.65	53.89±14.29	35.80±6.92	**36.24**±3.23	22.50±6.17	57.58±13.10	41.87±10.38	52.23±8.96
	G2P2	61.80±5.87	-	49.93±7.69	**41.28**±6.61	-	-	68.31±10.69	**46.47**±9.56	-
	GRAPHCLIP	61.79±4.05	-	40.45±11.11	36.08±2.89	-	-	41.26±25.34	38.55±15.13	-
	GFT	61.48±5.63	6.07±0.36	**62.44**±24.41	35.64±5.29	33.69±2.79	24.97±2.98	52.07±8.23	41.72±7.39	53.43±4.77
	UNIGRAPH2	**62.85**±5.69	4.57±0.96	53.98±17.78	31.79±6.16	29.39±2.53	**31.98**±2.80	**91.58**±8.90	37.94±10.85	50.22±2.48
Social	GCOPE	51.35±5.87	0.31±0.04	51.28±17.69	31.84±5.44	25.44±1.72	OOT	36.62±16.88	OOT	47.18±10.88
	MDGPT	55.74±6.15	**11.86**±0.69	52.89±14.94	40.60±7.21	34.40±3.25	25.04±6.46	61.36±21.16	38.99±8.98	**54.71**±10.57
	MDGFM	44.85±6.53	OOM	53.89±26.93	21.84±5.21	28.16±2.97	5.79±2.71	52.40±19.78	39.21±15.00	50.21±6.72
	SAMGPT	61.46±5.59	11.52±0.70	56.81±16.46	37.66±6.33	**36.14**±2.67	28.16±6.17	59.18±8.73	43.57±9.16	54.36±9.86
	G2P2	**63.44**±4.86	-	54.75±10.36	**41.97**±6.63	-	-	69.07±10.80	**46.74**±9.85	-
	GRAPHCLIP	40.18±2.96	-	36.40±11.17	19.57±4.46	-	-	54.22±32.89	36.94±20.91	-
	GFT	61.93±4.96	6.62±0.39	**60.14**±14.64	38.69±5.44	33.49±2.65	24.89±2.75	71.34±12.80	41.66±6.91	53.36±5.05
	UNIGRAPH2	60.91±6.72	3.60±0.93	52.22±18.52	33.88±6.82	31.95±3.20	**30.44**±3.42	**95.39**±8.87	38.74±9.38	50.43±3.74
Molecular	GCOPE	48.13±5.74	0.23±0.05	41.20±28.99	21.44±10.93	23.42±1.61	OOT	36.57±11.05	OOT	47.61±13.78
	MDGPT	58.18±7.08	9.78±0.80	55.71±13.79	40.07±7.14	34.74±3.38	**28.76**±8.15	62.79±20.83	38.25±8.81	**54.23**±10.93
	MDGFM	48.78±6.77	OOM	42.89±28.19	22.74±7.12	29.01±3.93	7.36±3.67	55.73±17.21	38.48±22.36	50.98±7.71
	SAMGPT	61.39±5.91	**11.07**±0.67	51.65±14.12	34.03±6.18	**36.44**±3.09	28.01±7.10	61.49±10.00	43.12±9.83	53.35±9.43
	G2P2	61.33±5.28	-	55.55±12.17	**40.92**±7.41	-	-	69.77±10.21	**46.93**±9.85	-
	GRAPHCLIP	34.47±2.57	-	36.45±11.50	18.20±6.39	-	-	50.63±14.36	37.67±18.93	-
	GFT	**61.70**±5.63	6.46±0.36	**60.13**±15.70	38.69±5.05	33.63±2.66	24.73±2.44	71.57±11.54	41.85±7.35	53.34±4.77
	UNIGRAPH2	57.53±5.98	2.62±0.62	49.56±23.19	27.18±6.35	30.28±2.32	17.24±2.95	**79.23**±8.18	39.12±11.24	50.99±5.83

Table 31: Complete Macro-F1 of 5-shot node classification on topic domain adaptation under three single-topic pre-training groups.

Pre-train	Methods	Citation		Social & Web			E-commerce	Finance		Proteins
		Pubmed	ogbn-mag	Wikipedia	Actor	Chameleon	Products	T-Finance	DGraph	ogbn-proteins
Citation	GCOPE	53.83±5.50	0.13±0.01	35.59±8.92	23.33±3.60	24.03±1.76	OOT	26.49±11.77	OOT	34.76±7.94
	MDGPT	53.22±8.58	5.72±0.83	35.59±10.76	31.39±4.46	32.54±3.24	8.91±1.68	44.41±12.02	27.72±5.64	50.96±9.83
	MDGFM	41.34±6.05	OOM	27.79±13.62	18.20±4.50	25.84±3.72	3.23±0.73	42.35±9.73	27.53±6.40	44.78±6.38
	SAMGPT	59.81±6.00	**9.40**±0.31	36.67±6.31	28.50±4.17	**35.74**±3.24	11.57±1.81	43.61±7.99	30.24±5.02	50.57±9.26
	G2P2	61.47±5.88	-	23.60±2.41	**33.37**±4.00	-	-	49.79±6.65	26.12±3.35	-
	GRAPHCLIP	61.68±4.03	-	20.79±4.86	31.66±1.92	-	-	29.71±14.61	22.70±6.30	-
	GFT	60.81±5.87	5.11±0.21	**41.39**±6.12	29.12±3.72	32.55±2.85	13.66±0.93	40.51±4.76	**30.98**±3.85	**51.43**±3.80
	UNIGRAPH2	**62.38**±5.73	2.09±0.09	36.24±7.97	25.33±4.06	26.96±2.83	**15.87**±1.04	**76.86**±12.64	27.22±4.31	48.56±1.49
Social	GCOPE	50.07±5.61	0.13±0.01	35.45±8.74	25.11±3.16	24.20±1.45	OOT	28.85±10.00	OOT	38.34±8.39
	MDGPT	55.53±6.74	**9.54**±0.44	36.13±6.76	**33.11**±4.36	32.77±3.88	13.36±1.83	45.13±13.40	28.63±4.38	**52.40**±9.76
	MDGFM	43.33±6.42	OOM	34.83±12.73	16.47±2.96	26.90±3.10	2.82±0.76	39.32±11.94	27.86±7.97	46.92±5.37
	SAMGPT	61.11±5.69	9.48±0.39	37.66±7.34	30.30±3.98	**35.62**±2.77	14.61±1.68	44.43±4.96	**31.15**±4.42	52.05±9.89
	G2P2	**63.08**±4.93	-	33.23±7.42	33.10±3.84	-	-	50.18±6.68	30.87±6.08	-
	GRAPHCLIP	39.13±3.37	-	18.93±4.15	16.01±2.38	-	-	34.28±16.47	22.53±10.16	-
	GFT	61.28±5.18	5.54±0.27	**40.23**±7.21	31.59±3.64	32.32±2.78	13.75±0.88	53.27±9.14	30.99±3.63	51.31±4.25
	UNIGRAPH2	60.89±6.59	1.42±0.08	35.46±8.49	26.86±4.51	29.42±3.46	**15.65**±1.02	**85.22**±10.81	27.64±4.38	48.44±2.26
Molecular	GCOPE	46.60±5.64	0.05±0.01	28.05±16.35	13.14±4.88	21.45±1.98	OOT	29.52±6.57	OOT	32.47±7.02
	MDGPT	58.18±7.32	7.70±0.51	37.47±6.31	**32.89**±4.90	33.03±3.44	13.12±2.06	46.17±13.03	28.20±4.31	51.26±9.85
	MDGFM	47.77±7.64	OOM	29.04±14.19	16.71±4.13	27.30±4.28	3.14±1.10	41.94±10.22	26.55±11.38	47.07±5.79
	SAMGPT	**61.15**±6.02	**8.99**±0.42	35.59±6.46	27.17±3.91	**35.71**±3.41	13.03±2.19	45.74±5.84	30.74±4.69	**52.26**±9.54
	G2P2	60.77±5.34	-	36.30±6.31	32.58±4.70	-	-	50.64±6.32	**32.60**±4.38	-
	GRAPHCLIP	32.45±2.23	-	19.15±5.06	14.00±3.17	-	-	36.59±6.89	21.86±8.54	-
	GFT	61.06±5.65	5.45±0.22	**40.22**±7.60	31.54±3.19	32.50±2.83	**13.63**±0.74	53.20±8.24	31.12±3.90	51.33±4.07
	UNIGRAPH2	57.36±5.89	0.88±0.06	33.42±10.91	21.98±4.48	28.40±2.87	9.05±0.71	**65.06**±5.02	27.90±4.02	49.24±3.51

Table 32: Complete evaluation of 5-shot edge classification and graph classification on topic domain adaptation under three single-topic pre-training groups.

Pre-train	Methods	Finance DGraph Acc	Finance DGraph MacroF	Common sense Wiki Acc	Common sense Wiki MacroF	Common sense WN18RR Acc	Common sense WN18RR MacroF	Molecule PCBA Acc	Molecule PCBA MacroF	Molecule BZR Acc	Molecule BZR MacroF
Citation	GCOPE	OOT	OOT	OOT	OOT	19.10±4.50	15.29±2.35	**65.12**±26.97	**37.70**±12.37	59.54±11.50	**53.19**±7.85
	MDGPT	10.14±2.03	7.40±0.88	11.63±2.85	3.10±0.30	19.36±3.03	16.12±1.37	56.62±14.60	35.76±6.33	55.33±10.14	50.80±7.72
	MDGFM	9.89±2.52	6.73±0.83	6.05±2.78	1.74±0.53	16.68±3.92	12.38±2.06	51.77±27.20	31.95±13.45	40.38±24.45	28.44±11.89
	SAMGPT	10.37±1.65	7.77±0.77	14.26±1.76	**5.41**±0.20	17.76±3.88	14.82±1.98	57.69±14.25	36.26±6.04	58.72±10.90	51.18±7.57
	G2P2	-	-	-	-	**20.18**±3.98	15.32±1.69	-	-	55.57±9.29	49.02±5.89
	GRAPHCLIP	-	-	-	-	17.52±1.97	**16.67**±1.30	-	-	**64.72**±13.04	50.29±6.15
	GFT	**11.28**±1.46	**8.55**±0.82	OOT	OOT	16.42±2.86	11.49±1.36	53.10±19.52	32.97±9.06	57.06±22.77	40.90±14.19
	UNIGRAPH2	10.05±2.37	6.82±0.92	**14.64**±3.07	5.35±0.96	15.68±3.98	10.56±1.65	50.68±21.04	32.53±7.62	54.46±14.57	45.68±7.98
Social	GCOPE	OOT	OOT	OOT	OOT	17.81±3.84	**15.36**±1.62	57.39±16.72	35.91±7.39	57.99±9.66	**51.66**±5.97
	MDGPT	10.82±2.62	7.70±1.32	13.34±1.99	4.86±0.26	16.60±2.63	13.78±1.46	56.88±13.91	35.94±5.99	54.08±10.18	48.87±8.00
	MDGFM	9.45±3.17	6.22±1.03	4.82±1.86	1.08±0.18	15.88±2.54	11.81±1.29	**60.17**±27.98	35.54±13.06	43.83±21.60	32.58±9.17
	SAMGPT	10.39±1.75	7.82±0.82	**15.04**±2.59	**5.75**±0.27	16.97±3.27	13.76±1.82	57.98±13.97	**36.41**±5.95	59.83±10.65	51.60±7.71
	G2P2	-	-	-	-	18.52±3.87	14.22±1.77	-	-	55.05±8.61	47.73±5.07
	GRAPHCLIP	-	-	-	-	**19.55**±4.77	13.52±1.80	-	-	**64.99**±14.17	47.35±7.86
	GFT	**11.25**±1.69	**8.49**±0.81	OOT	OOT	13.58±2.67	10.92±1.67	53.26±19.73	33.64±9.33	53.15±25.28	37.67±15.59
	UNIGRAPH2	10.13±2.84	6.82±0.91	14.98±3.05	5.10±0.32	13.79±3.30	10.09±1.39	52.16±21.40	32.13±10.15	53.58±16.09	45.60±10.28
Molecular	GCOPE	OOT	OOT	OOT	OOT	18.59±5.19	14.33±2.54	39.10±33.90	23.97±17.84	**67.41**±17.66	**55.30**±12.92
	MDGPT	10.76±2.58	7.60±1.02	12.46±2.07	3.89±0.36	16.09±1.90	14.03±1.43	56.94±13.67	35.99±5.96	52.45±10.12	48.70±8.69
	MDGFM	10.10±2.85	6.88±1.12	5.17±2.08	1.02±0.21	15.48±2.73	11.87±1.38	**58.68**±24.45	35.53±11.16	47.24±22.31	35.18±10.48
	SAMGPT	10.37±1.52	7.63±0.72	**14.07**±3.26	**5.19**±0.72	**18.80**±4.34	**15.09**±2.37	58.27±13.84	**36.54**±5.84	59.78±9.58	52.40±6.98
	G2P2	-	-	-	-	17.88±6.08	13.67±3.33	-	-	54.25±17.78	43.89±11.10
	GRAPHCLIP	-	-	-	-	16.15±4.63	10.67±1.65	-	-	59.92±9.91	50.78±5.59
	GFT	**11.10**±1.12	**8.45**±0.69	OOT	OOT	13.21±2.65	7.65±1.67	51.41±19.70	32.33±9.15	56.95±19.80	42.91±12.42
	UNIGRAPH2	9.87±2.75	6.66±0.99	12.08±2.38	3.53±1.05	13.26±4.17	7.43±1.75	51.57±17.75	32.94±9.13	54.51±7.51	47.14±4.98



(a) JSD between degree distributions.

(b) TF–IDF cosine similarity between node text.

Figure 5: Dataset-level structural and textual similarity across benchmark datasets.

C.4 DATASET-LEVEL SIMILARITY ANALYSIS

To further interpret the topic-domain results in Setting III, Fig. 5 compares graph-structure and node-text similarity across datasets. Structural similarity is measured by JSD between degree histograms with a shared bin range (Hartle et al., 2020) (lower is more similar), while textual similarity is measured by cosine similarity between dataset-level TF–IDF vectors over the 5,000 most frequent tokens (higher is more similar) and is reported only for text-attributed datasets.

Topic labels only weakly reflect dataset similarity. Although the mean JSD is lower within topics than across topics (0.23 vs. 0.33), the two distributions overlap substantially. For example, *DGraph* and *T-Finance* share the *Finance* label but have a JSD of 0.66, whereas the cross-topic pair *T-Finance* and *ogbn-proteins* has a JSD of 0.11. Textual similarity shows the same pattern: *Products* is more similar to *ogbn-arxiv* (0.91) than to *Amazon-HeTGB* (0.23), although only the latter shares its *E-commerce* label.

Table 33: Evaluation of one-shot edge classification and graph classification on format domain adaptation.

Methods	Relational (Edge Classification)				Textual (Graph Classif.)	
	Wiki		WN18RR		PCBA	
	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE	OOT	OOT	**17.76**±6.44	**12.87**±3.03	50.93±34.41	30.23±16.26
MDGPT	10.07±3.05	2.75±0.18	14.71±6.07	10.55±2.46	55.25±24.41	34.04±11.43
MDGFM	3.85±2.53	0.37±0.09	13.77±5.17	9.93±2.07	51.68±48.73	26.37±24.37
SAMGPT	**12.33**±3.59	**3.69**±0.18	13.53±4.88	9.97±2.11	**56.06**±22.56	34.63±10.44
G2P2	-	-	16.35±8.67	9.50±2.37	-	-
GRAPHCLIP	-	-	16.85±4.74	12.19±1.96	-	-
GFT	OOT	OOT	13.24±6.67	8.84±1.60	51.30±28.10	**35.62**±15.45
UNIGRAPH2	5.02±3.20	2.29±0.19	10.87±8.36	8.08±1.58	49.37±22.55	32.39±11.87

Table 34: Evaluation of 1-shot node classification on format domain adaptation under the FAGCN-backbone variant.

Methods	Heterogeneous		Heterophilic				Dynamic		Textual			
	ogbn-mag		Chameleon		Actor		Wikipedia		Pubmed		Products	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE (GCN)	0.29±0.01	0.18±0.01	22.32±1.86	18.99±2.51	20.07±7.27	14.33±3.34	30.65±27.78	21.64±15.09	35.51±6.93	28.92±5.84	OOT	OOT
MDGPT	2.31±0.65	1.24±0.29	27.62±4.41	23.83±4.36	25.85±6.33	20.59±4.38	47.73±17.60	33.40±8.38	45.67±9.63	43.70±9.62	15.06±5.20	8.02±1.27
MDGFM	1.86±0.40	1.40±0.23	26.55±4.10	24.75±3.59	**26.26**±4.22	**21.89**±3.25	50.18±3.72	**35.62**±1.73	43.71±6.98	43.29±7.30	9.91±1.85	6.13±0.97
SAMGPT	**4.89**±0.98	**3.28**±0.42	29.21±4.39	25.26±4.18	25.28±8.50	18.56±4.38	51.11±23.77	34.00±11.79	**48.45**±9.24	**45.31**±10.45	**20.76**±5.28	**10.55**±1.36
G2P2	-	-	-	-	21.36±10.99	13.18±4.70	**53.36**±30.18	31.81±14.98	39.52±5.92	38.40±5.71	-	-
GRAPHCLIP	-	-	-	-	19.22±18.62	6.76±5.42	10.02±16.93	7.98±9.12	37.86±3.73	20.81±3.48	-	-
GFT	0.54±0.38	0.17±0.07	25.89±3.09	20.11±3.92	24.71±6.12	19.29±3.24	49.27±8.96	34.85±4.26	39.31±5.98	35.95±7.47	5.01±2.02	2.86±0.61
UNIGRAPH2	3.67±0.34	2.58±0.13	**31.05**±4.54	**26.88**±4.68	25.08±7.67	19.11±4.57	48.36±21.56	33.18±10.53	43.05±8.91	41.70±9.38	18.36±3.87	10.00±1.29

These observations are consistent with the modest and inconsistent benefit of topic alignment in Setting III (Sec. 5.1). Because datasets assigned to the same topic can remain distant in structure or node-text distributions, dataset-level statistics can complement coarse topic labels when characterizing pre-training corpora and interpreting transfer behavior.

C.5 FORMAT DOMAIN ADAPTATION

The additional results for Setting IV, discussed in Sec. 5.1, are reported below. Specifically, we present the 1-shot edge and graph classification results in Table 33, the 5-shot node classification results in Table 38, and the 5-shot edge and graph classification results in Table 39. We additionally report the FAGCN-backbone variant in Tables 34, 35, 36, and 37. For most GFMs, the original homophilic backbone is replaced with FAGCN; GCOPE instead uses GCN as its homophilic-backbone counterpart because its default backbone is already FAGCN.

Table 35: Evaluation of 5-shot node classification on format domain adaptation under the FAGCN-backbone variant.

Methods	Heterogeneous		Heterophilic				Dynamic		Textual			
	ogbn-mag		Chameleon		Actor		Wikipedia		Pubmed		Products	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE (GCN)	0.29±0.01	0.18±0.01	23.98±1.62	23.33±1.66	29.10±6.56	21.09±2.91	43.15±19.04	30.90±9.43	51.69±4.69	50.44±4.65	OOT	OOT
MDGPT	7.34±1.05	5.05±0.59	31.72±2.53	29.72±2.83	33.48±5.72	26.84±4.06	46.96±13.59	33.46±6.44	58.47±7.76	57.68±8.12	26.47±3.71	12.38±1.00
MDGFM	OOM	OOM	31.35±2.85	29.44±2.66	32.46±4.68	26.89±3.54	49.79±3.13	35.35±1.52	51.66±7.20	51.57±7.52	29.60±3.47	14.34±0.91
SAMGPT	**11.24**±1.07	**8.69**±0.77	35.23±2.01	33.23±2.39	36.62±8.28	**28.99**±5.04	49.76±11.44	34.98±5.44	**60.56**±6.13	**60.23**±6.27	29.72±6.69	13.64±1.95
G2P2	-	-	-	-	26.04±10.86	16.21±5.21	50.93±34.98	31.34±17.01	47.44±5.62	46.24±5.37	-	-
GRAPHCLIP	-	-	-	-	**48.88**±0.00	13.13±0.00	31.48±37.56	18.55±17.67	38.56±1.68	20.16±2.37	-	-
GFT	0.89±0.55	0.40±0.28	30.76±2.59	28.45±3.27	33.67±4.88	26.78±3.36	**53.84**±13.89	**36.61**±5.81	51.73±4.46	51.04±4.63	13.45±2.37	8.16±0.72
UNIGRAPH2	6.43±0.50	4.52±0.16	**35.82**±3.05	**33.63**±3.26	27.83±10.02	20.57±5.86	51.22±24.94	33.93±11.87	58.80±5.13	58.65±5.26	**32.17**±2.76	**15.57**±0.95

Table 36: Evaluation of 1-shot edge classification and graph classification on format domain adaptation under the FAGCN-backbone variant.

Methods	Relational				Textual	
	Wiki		WN18RR		PCBA	
	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE (GCN)	OOT	OOT	9.37±3.32	7.07±1.61	42.50±30.44	26.62±15.70
MDGPT	7.30±3.37	1.46±0.14	12.74±4.14	9.82±1.93	**53.98**±22.52	**33.75**±10.26
MDGFM	3.66±3.23	0.69±0.12	9.58±1.16	6.71±0.81	53.91±33.33	31.84±16.14
SAMGPT	**9.75**±3.15	**2.94**±0.17	**16.44**±5.70	**11.06**±2.58	53.06±25.65	32.89±12.02
G2P2	-	-	7.47±9.32	3.56±2.99	-	-
GRAPHCLIP	-	-	14.64±13.39	8.07±3.42	-	-
GFT	OOT	OOT	10.58±3.77	7.11±1.94	47.16±29.84	30.93±18.79
UNIGRAPH2	3.95±3.17	1.01±0.14	8.53±5.51	4.70±1.40	49.43±21.27	31.61±10.22

Table 37: Evaluation of 5-shot edge classification and graph classification on format domain adaptation under the FAGCN-backbone variant.

Methods	Relational				Textual	
	Wiki		WN18RR		PCBA	
	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE (GCN)	OOT	OOT	16.49±4.14	**14.71**±1.86	49.59±16.32	32.52±7.60
MDGPT	11.70±2.79	3.20±0.25	15.78±3.30	13.42±1.72	55.55±15.93	35.18±7.05
MDGFM	**13.72**±4.01	4.94±0.34	11.16±1.30	9.10±1.05	52.72±25.14	32.90±12.67
SAMGPT	13.25±2.25	**5.08**±0.19	18.79±4.74	14.21±2.12	**59.16**±15.56	**36.76**±6.57
G2P2	-	-	10.17±10.70	4.98±3.32	-	-
GRAPHCLIP	-	-	**27.55**±14.17	10.62±2.53	-	-
GFT	OOT	OOT	11.78±3.10	8.65±1.43	52.62±20.39	33.74±8.79
UNIGRAPH2	6.42±2.10	3.59±0.13	10.47±4.46	6.57±1.66	51.58±14.92	32.96±6.95

Table 38: Evaluation of 5-shot node classification on format domain adaptation.

Methods	Heterogeneous		Heterophilic				Dynamic		Textual			
	ogbn-mag		Chameleon		Actor		Wikipedia		Pubmed		Products	
	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE	0.34±0.05	0.11±0.01	26.77±1.82	25.24±1.90	27.24±7.28	23.32±4.79	57.89±16.82	38.58±7.89	56.56±5.41	55.48±5.32	OOT	OOT
MDGPT	11.33±0.64	8.63±0.37	34.58±3.58	32.84±4.24	38.14±7.42	30.52±4.47	49.72±15.67	34.65±7.57	59.52±5.76	59.83±6.03	26.69±7.85	12.29±2.04
MDGFM	OOM	OOM	25.50±3.38	23.78±3.82	19.75±5.24	14.87±2.27	49.71±23.54	33.45±11.56	37.95±5.56	34.91±5.16	5.39±3.62	2.42±0.71
SAMGPT	**12.33**±0.71	**9.86**±0.33	**37.27**±3.48	**36.35**±3.50	**38.52**±5.37	**30.69**±3.36	52.67±12.71	36.30±5.89	63.88±5.30	63.78±5.29	**32.07**±8.02	**15.53**±1.98
G2P2	-	-	-	-	35.73±9.31	27.23±5.76	49.17±19.45	32.17±10.53	54.59±5.87	52.58±5.88	-	-
GRAPHCLIP	-	-	-	-	22.17±6.58	16.39±2.94	38.17±13.34	19.50±4.85	35.64±3.81	33.78±3.17	-	-
GFT	3.21±1.05	0.62±0.04	34.93±3.49	33.80±3.74	35.98±5.39	29.90±3.47	**63.23**±16.19	**41.26**±7.47	63.77±5.56	63.31±5.31	13.28±5.31	7.55±1.19
UNIGRAPH2	3.98±0.86	0.92±0.59	26.53±4.44	23.94±4.04	23.12±9.68	15.47±4.14	48.83±16.44	31.59±12.50	**74.36**±10.53	**68.29**±2.38	16.13±6.69	9.19±2.11

Table 39: Evaluation of 5-shot edge classification and graph classification on format domain adaptation.

Methods	Relational				Textual	
	Wiki		WN18RR		PCBA	
	Acc	MacroF	Acc	MacroF	Acc	MacroF
GCOPE	OOT	OOT	18.91±4.13	14.92±2.17	56.03±26.67	33.97±12.64
MDGPT	12.24±2.74	4.14±0.65	19.02±3.24	**15.73**±2.10	**56.56**±18.54	35.36±8.28
MDGFM	8.42±3.77	1.49±0.30	19.09±5.56	13.34±2.32	53.99±44.41	28.92±22.20
SAMGPT	**15.15**±2.81	**6.10**±0.22	17.06±3.39	14.08±1.74	56.29±14.33	35.66±6.18
G2P2	-	-	16.99±8.41	10.82±2.67	-	-
GRAPHCLIP	-	-	**20.68**±5.30	15.06±2.01	-	-
GFT	OOT	OOT	16.49±5.60	10.06±1.14	55.21±17.11	**40.30**±9.57
UNIGRAPH2	6.60±3.01	3.61±0.29	11.93±7.96	8.49±1.54	52.79±15.63	34.30±8.02

D AGGREGATE ANALYSIS PROCEDURE

D.1 POOLING PROCEDURE FOR FIGURES 1 AND 2

For both aggregate figures, we pool all available one- and five-shot Accuracy results across all downstream topic domains. Macro-F1 is not included. Missing or failed results, including '-', OOT, and OOM entries, are excluded rather than imputed. Each remaining model–dataset–shot comparison receives equal weight.

The aggregate analysis focuses on SAMGPT, G2P2, GFT, and UNIGRAPH2, which are generally competitive and comparatively stable over their applicable datasets. The remaining models are excluded only from these visual summaries because of pronounced dataset-specific limitations: GCOPE is OOT on *Products*, *DGraph*, and *Wiki*; MDGPT has high variance on *Wikipedia* and *T-Finance*; MDGFM degrades sharply on *Products* and *Wiki*; and GRAPHCLIP shares the same applicability as G2P2 but exhibits less stable performance across datasets and pre-training groups, often with large variance.

Let $A_{p,m,d,s}$ denote the mean accuracy under pre-training setting p, model m, downstream dataset d, and shot setting $s \in \{1, 5\}$. For Fig. 1, let $q(t)$ denote the topic-matched Setting III variant for downstream topic t: Setting III-C for citation, Setting III-S for social, and Setting III-M for molecular datasets. Each heatmap cell is computed as

$$H_{p,t} = \frac{100}{|\Omega_{p,t}|} \sum_{(m,d,s)\in\Omega_{p,t}} \frac{A_{p,m,d,s}}{A_{q(t),m,d,s}}, \tag{3}$$

where d belongs to topic t, and $\Omega_{p,t}$ contains only comparisons for which both the numerator and matched-topic denominator are available. Consequently, the matched Setting III diagonal is normalized to 100%. The figure reports $H_{p,t}$ for Setting I and the three Setting III variants over the three topics with a corresponding matched pre-training group.

For Fig. 2(a), the overperformance rate of Setting III variant p on downstream topic t is

$$O^{\mathrm{III}}_{p,t} = \frac{100}{|\Gamma_{p,t}|} \sum_{(m,d,s)\in\Gamma_{p,t}} \mathbb{I}[A_{p,m,d,s} > A_{\mathrm{I},m,d,s}], \tag{4}$$

where $\Gamma_{p,t}$ includes all valid comparisons from datasets assigned to topic t. All seven downstream topic domains are selected. For Fig. 2(b), the Setting IV rate for format domain f is

$$O^{\mathrm{IV}}_{f} = \frac{100}{|\Gamma_{f}|} \sum_{(m,d,s)\in\Gamma_{f}} \mathbb{I}[A_{\mathrm{IV},m,d,s} > A_{\mathrm{I},m,d,s}], \tag{5}$$

where Γ_f contains all valid comparisons for downstream datasets exhibiting format f; a dataset contributes to every applicable format category. The indicator uses a strict comparison, so a tie is not counted as an overperformance. Thus, the bars pool wins over all selected models, datasets, and shots without first averaging within a model or dataset.