BENCHMARKING RELIABILITY IN GRAPH REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph representation learning (GRL) has evolved from topology-only graph embeddings to task-specific supervised GNNs and, more recently, to reusable representations and graph foundation models (GFMs). However, existing GRL benchmarks primarily evaluate clean-setting performance, transfer, adaptation, and task coverage, while reliability-relevant evaluations are typically conducted one axis at a time under different method pools and protocols. As a result, it remains unclear how modern GRL methods compare under matched conditions when deployment conditions affect graph signals, graph contexts, label support, structural groups, or predictive evidence. We introduce GRL-Reliability, a multi-axis benchmark for evaluating GRL reliability. GRL-Reliability evaluates twelve representative methods, spanning topology-only embeddings, supervised GNNs, self-supervised graph models, and GFMs, on twenty-five graph datasets under standardized evaluation conditions while preserving method-native adaptation where available. The benchmark covers five evaluation axes: corruption robustness, OOD generalization, class imbalance, fairness, and interpretation, with per-axis and subcondition reporting rather than a single aggregate score. Our analysis reveals three cross-axis insights. First, even tests that intervene on the same graph factor can produce distinct method profiles, showing that closely related tests capture different aspects of reliability. Second, foundation-era methods show axis-specific strengths rather than broad reliability advantages. Third, degree shift and rare-class coverage remain challenging across the evaluated method set, revealing deployment regimes where current GRL methods still have substantial room for improvement. The benchmark, evaluation protocols, and code are available at: `https://anonymous.4open.science/r/GRL-Reliability-1889/`.

1 INTRODUCTION

Foundation models (FMs) (Bommasani et al., 2021) have shifted machine learning in natural language processing (Devlin et al., 2019) and computer vision (Radford et al., 2021) from task-specific training toward transferable representations that can be adapted across tasks and domains. Graph learning (Wu et al., 2020), which underlies applications including molecular design (Stokes et al., 2020), recommendation systems (Ying et al., 2018), and financial networks (Dou et al., 2020), is undergoing a parallel shift. In particular, graph representation learning (GRL) has progressed from topology-only graph embeddings (Perozzi et al., 2014; Grover & Leskovec, 2016) and task-specific supervised graph neural networks (GNNs) (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017) toward transferable graph representations, including self-supervised pretrained models (Thakoor et al., 2022; Hou et al., 2022) and graph foundation models (GFMs) (Wang et al., 2025b), with the goal of generalizing across graphs, domains, and task formats.

As these representations are deployed in diverse downstream settings, deployment conditions expose reliability risks that clean-setting performance does not capture (Dai et al., 2024). These risks can arise through graph-specific channels, including corrupted attributes or edges (Zheng et al., 2021), distribution shifts across graph contexts (Gui et al., 2022), skewed label support (Ma et al., 2025), structural or demographic group disparity (Dong et al., 2024b), and unreliable predictive evidence (Agarwal et al., 2023). In financial graphs, for example, corrupted links or rare fraud labels can affect risk prediction in ways that clean aggregate performance may hide (Dou et al., 2020); in scientific or molecular graphs, unreliable attribution evidence can undermine the trustworthiness of

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.

model outputs (Amara et al., 2022). Throughout, we use *reliability* in an operational sense: whether a method's predictions remain accurate, balanced, and well-grounded when such risks arise. Together, these risks raise a broader question: *how do modern GRL methods behave across graph-specific deployment conditions under comparable evaluation settings?*

Existing evaluation efforts only partially answer this question. GRL and GFM benchmarks have made substantial progress in measuring clean-setting performance, transfer, adaptation, and task coverage across datasets, tasks, and domains (Hu et al., 2020; Xu et al., 2024; Chen et al., 2024b; Yu et al., 2026). In parallel, trustworthy GRL evaluation has developed axis-specific benchmarks and protocols for many of these concerns (Zheng et al., 2021; Gui et al., 2022; Ma et al., 2025; Dong et al., 2024b; Agarwal et al., 2023). However, these two lines remain disconnected: GRL benchmarks typically do not evaluate these axes, while such evaluation efforts are often organized one axis at a time, under narrower method families or different protocols. What remains missing is a unified, multi-axis reliability evaluation of the modern GRL spectrum under comparable conditions. Such a matched evaluation can determine whether related tests provide redundant or complementary evidence, whether foundation-era methods deliver broad reliability gains across deployment conditions, and which deployment regimes remain challenging across the evaluated method set. When only one risk matters, an axis-specific benchmark suffices; GRL-Reliability targets settings where several risks must be considered together.

Evaluating several risks together, however, is not simply a matter of combining existing axis-specific benchmarks. A unified evaluation must account for heterogeneity across evaluation axes, model interfaces, and graph settings. First, evaluation axes probe different properties of GRL systems, so their results should be reported by axis and subcondition rather than collapsed into a single score (Liang et al., 2023; Wang et al., 2023). Second, GRL methods differ in their input requirements and adaptation interfaces, requiring a standardized evaluation substrate while preserving method-native objectives and adaptation procedures (Xu et al., 2024; Chen et al., 2024b; Yu et al., 2026). Finally, because reliability risks can manifest differently across graph domains and task formats, the evaluation should span diverse datasets rather than a single graph family (Hu et al., 2020; Gui et al., 2022; Feng et al., 2024).

Guided by these requirements, we introduce GRL-Reliability, a multi-axis reliability benchmark for GRL. GRL-Reliability evaluates twelve representative GRL methods on twenty-five graph datasets spanning diverse domains and task formats. Under standardized evaluation conditions, feature-consuming methods use shared text-attribute embeddings, topology-only embedding methods use shared graph structure on compatible task formats, and methods with a pretraining stage share a common source corpus while preserving their original pretraining objectives and native downstream adaptation where available. Our contributions are:

- **Benchmark scope.** GRL-Reliability positions reliability evaluation at the level of GRL, covering topology-only graph embeddings, supervised GNNs, self-supervised GRL methods, and GFMs across diverse graph tasks and domains.
- **Multi-axis reliability protocol and evaluation.** GRL-Reliability evaluates five deployment-relevant evaluation axes: corruption robustness, OOD generalization, class imbalance, fairness, and interpretation. It reports per-axis and subcondition results rather than collapsing graph reliability into a single aggregate score.
- **Cross-axis empirical insights.** GRL-Reliability shows that even structurally related tests can expose distinct reliability properties, that foundation-era methods exhibit axis-specific rather than broad reliability advantages, and that degree shift and rare-class coverage remain challenging across the evaluated method set. These patterns remain qualitatively consistent on pretraining-unseen evaluation graphs.

2 Related Work

Graph representation learning. GRL has progressed from topology-only embeddings (Perozzi et al., 2014; Grover & Leskovec, 2016) and supervised GNNs (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017; Zhang et al., 2019) to self-supervised and transferable representation learning (Veličković et al., 2019; You et al., 2020; Ju et al., 2023; Thakoor et al., 2022; Hou et al., 2022; Tian et al., 2023; Wang et al., 2025d;e), and more recently to foundation-era models that unify

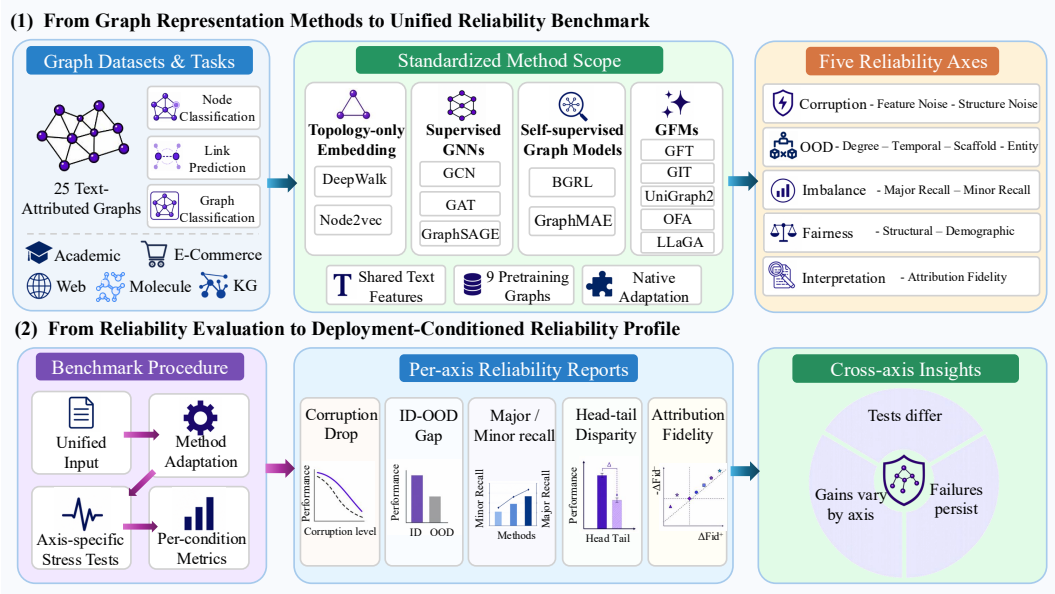


Figure 1: Overview of GRL-Reliability. We evaluate twelve GRL methods across five evaluation axes under standardized evaluation conditions.

tasks, transfer across graphs, or connect graph structure with language models (Liu et al., 2024; Wang et al., 2024a; 2025c; He et al., 2025; Yuan et al., 2024a; Guo et al., 2026; Chen et al., 2024a; Tang et al., 2024; Wang et al., 2025b). Existing benchmarks for this spectrum primarily evaluate clean performance, transfer, adaptation, and task coverage (Hu et al., 2020; Xu et al., 2024; Chen et al., 2024b; Yu et al., 2026).

Reliability and trustworthiness in GRL. Existing work studies corruption robustness (Zheng et al., 2021; Li et al., 2021; Wang et al., 2024b), OOD generalization (Gui et al., 2022; Wang et al., 2025a), class imbalance (Ma et al., 2025), fairness (Dong et al., 2024a;b), and explanation fidelity (Agarwal et al., 2023; Amara et al., 2022), alongside broader trustworthy graph learning studies (Dai et al., 2024; Wang et al., 2026; Yuan et al., 2024b). These evaluations are typically organized one axis at a time under different inputs, protocols, and method pools, and rarely cover the broader modern GRL spectrum, particularly recent GFMs. This fragmentation makes cross-axis behavior difficult to compare under a common evaluation setting. GRL-Reliability addresses this gap by evaluating methods from topology-only embeddings to GFMs across multiple evaluation axes under matched conditions.

3 GRL-Reliability Benchmark and Evaluation

3.1 Setup

GRL-Reliability evaluates GRL methods under graph-specific deployment conditions by standardizing inputs, splits, evaluation protocols, and metrics while preserving method-native objectives and adaptation where available. Shared inputs reduce variation from feature preprocessing, so the resulting comparisons characterize method behavior under this common evaluation setting rather than each method's best performance under alternative preprocessing pipelines. Results are reported by evaluation axis and subcondition rather than collapsed into a single aggregate leaderboard.

Method Scope. GRL-Reliability covers twelve methods spanning four method families. Topology-only baselines include DeepWalk (Perozzi et al., 2014) and Node2vec (Grover & Leskovec, 2016). Supervised message-passing baselines include GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), and GraphSAGE (Hamilton et al., 2017). Self-supervised graph representation methods include BGRL (Thakoor et al., 2022) and GraphMAE (Hou et al., 2022). Foundation-era graph

models include GFT (Wang et al., 2024a), GIT (Wang et al., 2025c), UniGraph2 (He et al., 2025), OFA (Liu et al., 2024), and LLaGA (Chen et al., 2024a).

Datasets and Protocol. GRL-Reliability evaluates twenty-five text-attributed graphs covering node classification, link prediction, and graph classification across seven domains; dataset statistics are reported in Appendix A. For feature-consuming methods, each node, edge, atom, or relation text field is embedded with Sentence-BERT (Reimers & Gurevych, 2019) into a shared 768-dimensional feature space. Topology-only methods use the same graph structure on compatible task formats. Pretrained methods use a common nine-graph pretraining corpus following GFT (Wang et al., 2024a), while supervised and topology-only baselines are fitted directly on each downstream graph. All methods use shared downstream splits, evaluation protocols, and metric definitions where applicable. Each evaluation axis is therefore evaluated on the subset of datasets whose task format, label structure, split metadata, or attribution interface supports the corresponding evaluation protocol; Appendix D summarizes this axis-specific applicability.

Pretraining Overlap. Several evaluation graphs also appear in the shared pretraining corpus. We therefore additionally analyze the sixteen pretraining-unseen datasets, excluding corpus members and, conservatively, arXiv-derived variants, with each evaluation axis evaluated on its applicable subset. The main qualitative findings remain consistent with those of the full benchmark (Appendix E).

Evaluation Axes. GRL-Reliability evaluates five representative evaluation axes: corruption robustness, OOD generalization, class imbalance, fairness, and interpretation. Corruption separates feature perturbation from edge deletion; OOD separates degree, temporal, scaffold, and inductive entity shifts; imbalance reports major- and minor-class recall; fairness reports structural and demographic group disparities; and interpretation measures whether attributed substructures support the prediction better than random evidence. Capability transfer, calibration, privacy, and adaptive adversarial robustness are outside the current scope.

3.2 Per-Axis Reliability Evaluation

The following subsections instantiate these five axes as concrete evaluation protocols, report results across compatible methods, and summarize the main empirical finding.

3.2.1 Corruption Robustness: Feature and Structure Noise

Corruption evaluates input-side stability under two noise channels: feature noise, which degrades node attributes, and structure noise, which removes observed edges. Feature noise applies to feature-consuming methods, while structure noise applies to methods relying on topology. This axis asks whether a graph representation remains useful when attribute signal or graph structure is corrupted.

Benchmark. We instantiate two corruption channels following common graph robustness stress tests (Liu et al., 2021; Rong et al., 2020). **Feature noise** adds Gaussian perturbations to node features, with severity controlled by relative scale $\sigma_{\text{rel}} \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$; it is inapplicable to topology-only embeddings. **Structure noise** (random edge deletion) removes edges, with severity controlled by deletion probability $p \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$; DeepWalk and Node2vec, which lack an inductive inference path, are refitted on each edge-deleted graph (Appendix D.1). Table 1 reports the strongest setting for each channel, σ_{rel}=2.0 for feature noise and p=0.5 for structure noise, on the shared node-classification datasets. We report perturbed accuracy and Drop = Clean − Perturbed; lower drop indicates stronger corruption stability. Full severity sweeps are provided in Appendix G.1.

Evidence. Table 1 shows that relative robustness depends on which graph signal is corrupted. LLaGA is stable under feature noise but degrades under edge deletion (2.00 vs. 5.44 drop), whereas UniGraph2 shows the reverse (5.23 vs. 1.42). Figure 2 traces these differences across severity levels, and the full sweeps in Appendix G.1 show the same signal dependence.

Finding 1. Corruption robustness is signal-specific rather than a single property. Robustness to noisy features does not imply robustness to missing edges, and vice versa, so robustness should be assessed for the graph signal expected to fail.

Table 1: **Sev5 corruption summary.** Accuracy (%) and clean-to-corrupted drop on four shared NC datasets. Feature noise uses Gaussian perturbation with σ=2.0; structure noise uses random edge deletion with p=0.5. Sev5 is the strongest severity; Drop = Clean − Sev5 (↓ is more robust). Cells show mean$_{\pm\text{std}}$ over 5 seeds; aggregate std is across datasets. **Bold**/underline: best/second per column. "–" indicates feature noise is inapplicable.

	Feature Noise (σ=2.0)						Structure Noise (p=0.5)					
Method	**Citeseer**	**DBLP**	**Pubmed**	**Wikics**	**Sev5**	**Drop↓**	**Citeseer**	**DBLP**	**Pubmed**	**Wikics**	**Sev5**	**Drop↓**
DeepWalk (Perozzi et al., 2014)	–	–	–	–	–	–	$31.7_{\pm1.7}$	$50.3_{\pm0.5}$	$52.0_{\pm1.6}$	$66.8_{\pm1.0}$	$50.2_{\pm14.4}$	$9.63_{\pm7.3}$
Node2vec (Grover & Leskovec, 2016)	–	–	–	–	–	–	$31.0_{\pm1.5}$	$48.3_{\pm1.3}$	$52.6_{\pm1.2}$	$66.7_{\pm0.8}$	$49.6_{\pm14.7}$	$9.84_{\pm6.5}$
GCN (Kipf & Welling, 2017)	$67.1_{\pm1.6}$	$\underline{73.3}_{\pm0.2}$	**77.4**$_{\pm2.1}$	$\underline{77.9}_{\pm0.7}$	$\underline{73.9}_{\pm5.0}$	**1.18**$_{\pm0.2}$	$66.8_{\pm0.8}$	$\underline{73.5}_{\pm0.3}$	**77.2**$_{\pm1.8}$	$76.9_{\pm0.7}$	$\underline{73.6}_{\pm4.8}$	$1.44_{\pm0.6}$
GAT (Veličković et al., 2018)	$\underline{67.3}_{\pm1.4}$	**73.7**$_{\pm0.4}$	$\underline{77.2}_{\pm0.9}$	$77.6_{\pm0.6}$	**74.0**$_{\pm4.8}$	$\underline{1.29}_{\pm0.4}$	**67.7**$_{\pm1.4}$	**74.0**$_{\pm0.4}$	$\underline{76.6}_{\pm1.0}$	$76.6_{\pm0.5}$	**73.7**$_{\pm4.2}$	$1.51_{\pm0.7}$
GraphSAGE (Hamilton et al., 2017)	**67.3**$_{\pm2.0}$	$70.4_{\pm1.8}$	$74.2_{\pm1.4}$	$75.5_{\pm1.2}$	$71.9_{\pm3.7}$	$2.41_{\pm0.7}$	$\underline{67.0}_{\pm0.5}$	$72.6_{\pm0.8}$	$74.8_{\pm2.0}$	$76.0_{\pm0.3}$	$72.6_{\pm4.0}$	$1.84_{\pm0.6}$
BGRL (Thakoor et al., 2022)	$63.9_{\pm0.7}$	$69.6_{\pm0.5}$	$72.0_{\pm2.3}$	**78.1**$_{\pm0.7}$	$70.9_{\pm5.9}$	$1.98_{\pm1.0}$	$64.8_{\pm0.7}$	$70.6_{\pm0.3}$	$73.4_{\pm2.7}$	**78.0**$_{\pm0.6}$	$71.7_{\pm5.5}$	**1.16**$_{\pm1.0}$
GraphMAE (Hou et al., 2022)	$63.8_{\pm0.8}$	$68.7_{\pm1.3}$	$76.2_{\pm0.7}$	$77.7_{\pm0.9}$	$71.6_{\pm6.5}$	$2.99_{\pm1.3}$	$65.5_{\pm0.9}$	$71.5_{\pm0.7}$	$76.4_{\pm0.8}$	$\underline{77.8}_{\pm0.5}$	$72.8_{\pm5.6}$	$1.83_{\pm0.7}$
GFT (Wang et al., 2024a)	$59.7_{\pm0.9}$	$58.5_{\pm1.6}$	$73.0_{\pm1.2}$	$67.2_{\pm1.3}$	$64.6_{\pm6.8}$	$7.36_{\pm3.3}$	$57.7_{\pm0.5}$	$63.9_{\pm1.2}$	$74.4_{\pm1.5}$	$75.6_{\pm0.8}$	$67.9_{\pm8.6}$	$3.85_{\pm3.7}$
GIT (Wang et al., 2025c)	$49.9_{\pm1.5}$	$71.5_{\pm0.1}$	$73.5_{\pm1.6}$	$63.0_{\pm2.0}$	$64.5_{\pm10.7}$	$9.47_{\pm8.2}$	$61.5_{\pm1.0}$	$72.8_{\pm0.2}$	$74.4_{\pm1.4}$	$75.8_{\pm0.7}$	$71.1_{\pm6.5}$	$3.00_{\pm2.3}$
UniGraph2 (He et al., 2025)	$60.4_{\pm0.7}$	$59.9_{\pm3.4}$	$71.4_{\pm2.0}$	$70.4_{\pm2.1}$	$65.5_{\pm6.2}$	$5.23_{\pm1.8}$	$64.2_{\pm1.0}$	$66.8_{\pm2.3}$	$72.6_{\pm1.7}$	$72.2_{\pm2.0}$	$68.9_{\pm4.1}$	$\underline{1.42}_{\pm1.7}$
OFA (Liu et al., 2024)	$50.8_{\pm1.6}$	$54.2_{\pm0.7}$	$62.3_{\pm3.0}$	$64.2_{\pm1.2}$	$57.9_{\pm6.4}$	$11.98_{\pm0.9}$	$53.5_{\pm2.3}$	$64.2_{\pm1.8}$	$69.1_{\pm3.4}$	$69.1_{\pm0.8}$	$64.0_{\pm7.3}$	$5.89_{\pm3.9}$
LLaGA (Chen et al., 2024a)	$61.5_{\pm0.7}$	$68.2_{\pm1.9}$	$71.9_{\pm2.8}$	$73.2_{\pm0.9}$	$68.7_{\pm5.2}$	$2.00_{\pm1.0}$	$53.5_{\pm2.2}$	$68.3_{\pm1.3}$	$67.1_{\pm2.6}$	$72.2_{\pm1.1}$	$65.3_{\pm8.1}$	$5.44_{\pm4.6}$

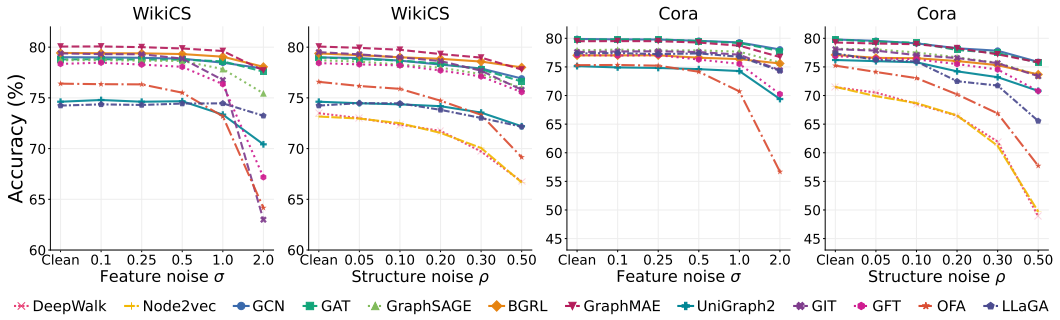


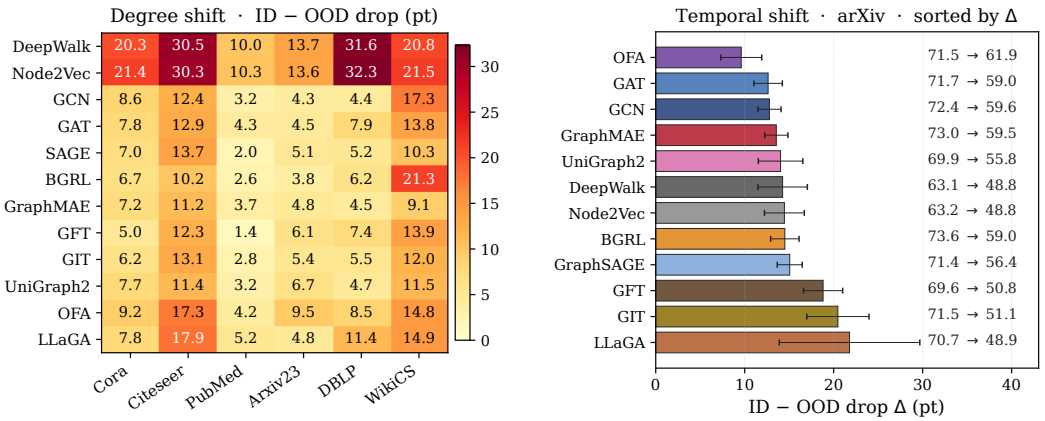
Figure 2: **Corruption severity trajectories.** Test accuracy under feature- and structure-noise sweeps on WikiCS and Cora. Curves denote methods; per-dataset curves in Appendix Figures 6–7.

3.2.2 OOD Generalization: Shifts Across Graph Contexts

OOD generalization evaluates whether graph representations remain useful when the graph context changes across structure, time, molecular scaffold, or entity set. We study four OOD shift mechanisms: degree shift, temporal shift, scaffold shift, and inductive-entity shift.

Benchmark. We instantiate four OOD shift mechanisms following mechanism-level evaluation protocols in graph OOD benchmarks (Gui et al., 2022). **Degree shift** separates high-degree training nodes from low-degree OOD nodes by degree quantile on 13 node-classification datasets. **Temporal shift** trains on arXiv papers with publication year $\leq$ 2010 and evaluates on papers from $\geq$ 2017. **Scaffold shift** compares random and Bemis–Murcko scaffold splits on BACE and Tox21 (Hu et al., 2020). **Inductive-entity shift** is operationalized as KG completion on WN18RR and FB15K237, where test-time candidate entities are held out from downstream KG training. We report ID-to-OOD accuracy drop for degree and temporal shifts, $\Delta\text{AUC} = \text{AUC}_{\text{rand}} - \text{AUC}_{\text{scaffold}}$ for scaffold shift, and MRR / Hits@10 for inductive-entity shift. Because arXiv, WN18RR, and FB15K237 belong to the shared pretraining corpus, temporal and inductive-entity results reflect behavior under this pretraining protocol; for KG completion, test entities are unseen in downstream training rather than in pretraining.

Evidence. Figure 3 and Table 2 show that OOD rankings change with the evaluated shift mechanism. Under degree shift across six node-classification datasets, topology-only embeddings have the largest average drops, while GraphMAE has the smallest average drop (6.75 points across the six datasets in Figure 3a). On the arXiv temporal split, OFA shows the smallest drop (9.61 points). Across the two scaffold-shift datasets evaluated, the method with the smallest gap differs: GraphSAGE shows the smallest BACE gap (9.7 ΔAUC), while UniGraph2 shows the smallest Tox21 gap (5.4 ΔAUC). Across the two inductive-entity KG datasets evaluated, the leader also changes: BGRL leads



(a) **Degree shift.** ID–OOD accuracy drop across 12 methods on 6 NC datasets. Full results are in Table 12.

(b) **Temporal shift.** ID–OOD accuracy drop on arXiv, sorted by drop size.

Figure 3: **Node-classification OOD shifts.** Larger values indicate weaker OOD generalization; darker cells and longer bars denote larger drops.

Table 2: **OOD shift mechanisms beyond node classification.** (a) **Scaffold shift** on BACE / Tox21: ROC-AUC under random versus Bemis–Murcko scaffold splits (80/10/10, OGB convention (Hu et al., 2020)); $\Delta = \mathrm{AUC}_{\mathrm{rand}} - \mathrm{AUC}_{\mathrm{scaf}}$, where larger Δ indicates a larger scaffold-generalization gap. (b) **Inductive-entity shift** on WN18RR / FB15K237, operationalized as KG completion with test-time candidate entities held out from downstream KG training; MRR / Hits@10, higher is better. **Bold**/underline: best/second per column. Per-cell values are mean$_{\pm\mathrm{std}}$ over 5 seeds; Δ is computed from unrounded means.

(a) Mol scaffold OOD (BACE / Tox21)

Method	BACE rand.	BACE scaf.	BACE Δ ↓	Tox21 rand.	Tox21 scaf.	Tox21 Δ ↓
GCN (Kipf & Welling, 2017)	64.5$_{\pm 0.4}$	52.1$_{\pm 0.6}$	12.3	78.5$_{\pm 0.2}$	72.5$_{\pm 0.2}$	5.9
GAT (Veličković et al., 2018)	62.9$_{\pm 0.3}$	48.1$_{\pm 1.1}$	14.7	78.6$_{\pm 0.4}$	71.2$_{\pm 0.3}$	7.4
GraphSAGE (Hamilton et al., 2017)	66.7$_{\pm 0.9}$	56.9$_{\pm 0.9}$	**9.7**	80.9$_{\pm 0.2}$	74.9$_{\pm 0.2}$	6.0
BGRL (Thakoor et al., 2022)	**85.3**$_{\pm 0.8}$	61.7$_{\pm 2.1}$	23.6	82.4$_{\pm 0.3}$	76.0$_{\pm 0.3}$	6.4
GraphMAE (Hou et al., 2022)	76.5$_{\pm 0.8}$	54.3$_{\pm 0.6}$	22.2	80.4$_{\pm 0.5}$	73.1$_{\pm 0.7}$	7.4
GFT (Wang et al., 2024a)	70.0$_{\pm 5.8}$	54.3$_{\pm 1.4}$	15.7	80.8$_{\pm 0.6}$	73.4$_{\pm 0.8}$	7.5
GIT (Wang et al., 2025c)	82.8$_{\pm 0.6}$	65.2$_{\pm 0.6}$	17.5	**83.0**$_{\pm 0.4}$	**76.0**$_{\pm 0.4}$	6.9
UniGraph2 (He et al., 2025)	68.5$_{\pm 2.1}$	56.7$_{\pm 2.8}$	11.7	75.5$_{\pm 1.6}$	70.2$_{\pm 0.9}$	**5.4**
OFA (Liu et al., 2024)	84.0$_{\pm 1.3}$	**66.6**$_{\pm 1.6}$	17.3	81.8$_{\pm 0.5}$	75.5$_{\pm 0.8}$	6.3
LLaGA (Chen et al., 2024a)	79.0$_{\pm 0.7}$	56.0$_{\pm 0.9}$	23.0	79.8$_{\pm 2.2}$	73.6$_{\pm 0.9}$	6.2

(b) Inductive KG (WN18RR / FB15K237)

Method	WN18RR MRR ↑	WN18RR Hits@10 ↑	FB15K237 MRR ↑	FB15K237 Hits@10 ↑
GCN (Kipf & Welling, 2017)	0.192$_{\pm .014}$	0.346$_{\pm .015}$	0.098$_{\pm .006}$	0.179$_{\pm .005}$
GAT (Veličković et al., 2018)	0.167$_{\pm .011}$	0.363$_{\pm .021}$	0.039$_{\pm .004}$	0.092$_{\pm .017}$
GraphSAGE (Hamilton et al., 2017)	0.134$_{\pm .027}$	0.262$_{\pm .031}$	0.172$_{\pm .011}$	0.378$_{\pm .015}$
BGRL (Thakoor et al., 2022)	**0.231**$_{\pm .009}$	0.443$_{\pm .027}$	0.088$_{\pm .004}$	0.153$_{\pm .002}$
GraphMAE (Hou et al., 2022)	0.224$_{\pm .014}$	**0.541**$_{\pm .025}$	0.121$_{\pm .009}$	0.256$_{\pm .015}$
GFT (Wang et al., 2024a)	0.157$_{\pm .005}$	0.339$_{\pm .013}$	0.174$_{\pm .005}$	0.364$_{\pm .009}$
GIT (Wang et al., 2025c)	0.094$_{\pm .013}$	0.212$_{\pm .010}$	0.161$_{\pm .005}$	0.346$_{\pm .010}$
UniGraph2 (He et al., 2025)	0.148$_{\pm .007}$	0.315$_{\pm .015}$	0.064$_{\pm .008}$	0.127$_{\pm .012}$
OFA (Liu et al., 2024)	0.075$_{\pm .011}$	0.139$_{\pm .013}$	**0.186**$_{\pm .013}$	**0.385**$_{\pm .019}$
LLaGA (Chen et al., 2024a)	0.152$_{\pm .019}$	0.270$_{\pm .027}$	0.082$_{\pm .012}$	0.155$_{\pm .019}$

WN18RR in MRR (0.231) but falls to 0.088 on FB15K237, whereas OFA leads FB15K237 (0.186) but drops to 0.075 on WN18RR. These results illustrate subcondition sensitivity rather than a single OOD ranking.

Finding 2. OOD reliability is application-shift specific. In the evaluated settings, degree shifts in node-classification graphs, the arXiv temporal split, scaffold splits in molecules, and inductive-entity shifts in KGs expose different representation behaviors. Model selection should therefore be conditioned on the anticipated deployment shift rather than on a single aggregate OOD score.

3.2.3 Class Imbalance: Major and Minor Recall

Class imbalance evaluates whether strong average performance reflects balanced class-wise recall or is driven mainly by majority-class predictions under skewed label support.

Benchmark. We measure class-imbalance behavior on node classification at imbalance ratio ρ=10. Classes are sorted by their original training-set sample count; the lower half are treated as minor classes and downsampled to n_{major}/ρ, where n_{major} is the largest training count among the upper-half major classes. Full protocol details are in Appendix D.3. Each method is adapted on the imbalanced training graph using its native pipeline. We report per-class recall on the test split,

averaged separately across major and minor classes. Figure 4 shows four representative datasets where all twelve methods are covered; full per-dataset results for eight categorical-label NC datasets and $\rho \in \{5, 10, 20\}$ are reported in Appendix Table 13.

Evidence. Figure 4 shows that major-class and minor-class recall separate under skewed label support. This separation is most visible for topology-only graph embeddings, which have near-zero minor-class recall on Cora and CiteSeer despite much higher major-class recall. Among feature-consuming methods, the strongest minor-class method changes by dataset: LLaGA leads on Cora, GAT on CiteSeer, GraphMAE on WikiCS, and UniGraph2 on Elec-Computers. Minor-class gains also need not align with majority-class recall: on Elec-Computers, UniGraph2 has the best minor recall but lower major recall than LLaGA (75.0% vs. 81.1%), whereas OFA keeps competitive major recall while dropping to 36.4% minor recall on WikiCS. Overall, class imbalance exposes a strong separation between majority- and minority-class coverage rather than a uniform accuracy degradation pattern.

Finding 3. Class imbalance creates a rare-class coverage failure in graph prediction. Under skewed label support, strong majority-class recall can hide severe minority-class collapse, while methods with better minority coverage may sacrifice majority performance. For graph tasks where rare classes carry the primary risk, model selection should prioritize minority coverage and balanced recall rather than average accuracy.

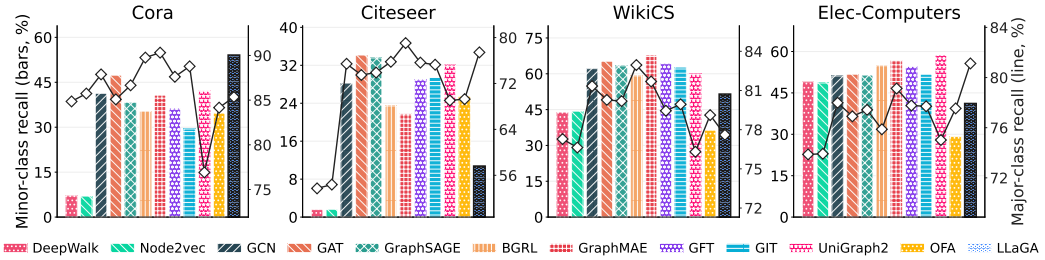


Figure 4: **Major- and minor-class recall under class imbalance.** At imbalance ratio ρ=10, bars show minor-class recall and lines show major-class recall on four representative node-classification datasets. High major-class recall does not necessarily imply high minor-class recall.

3.2.4 FAIRNESS: STRUCTURAL HEAD–TAIL DISPARITY

Fairness evaluates whether prediction quality differs across groups. In GRL, a natural group axis is structural position: high-degree nodes receive more neighborhood evidence, while low-degree nodes have thinner local context.

Benchmark. We evaluate structural fairness on node classification. For each dataset, test nodes are split into *head* nodes (top 20% by degree) and *tail* nodes (bottom 20% by degree), and we report the signed head-minus-tail accuracy gap $\Delta = \mathrm{Acc}_{\mathrm{head}} - \mathrm{Acc}_{\mathrm{tail}}$. A smaller $|\Delta|$ indicates more balanced head/tail performance; $\Delta > 0$ indicates head-favoring behavior, while $\Delta < 0$ indicates tail-favoring behavior. Each method is adapted with its native pipeline, and no fairness-aware training intervention is applied. Figure 5 shows four representative datasets where all twelve methods are covered. Full structural-fairness results on eight NC datasets and demographic-fairness results on Tolokers are reported in Appendix G.4; because Tolokers is the only dataset exposing a compatible sensitive attribute, the demographic analysis is an exploratory case study rather than a cross-dataset conclusion.

Evidence. Figure 5 shows that head–tail disparity is graph-conditioned rather than family-conditioned. Most methods favor head nodes on Cora, PubMed, WikiCS, and ElePhoto, but the size of this advantage varies substantially across graphs and methods. The topology-only graph embeddings often show large head-favoring gaps, indicating that purely structural representations can exhibit large degree-position disparities. The smallest absolute gap also rotates across datasets: UniGraph2 is smallest on Cora, BGRL on PubMed, OFA on WikiCS, and LLaGA on ElePhoto. Individual methods do not carry a fixed fairness profile: for example, GFT has a large gap on WikiCS ($+14.58$ pp) but a

much smaller gap on ElePhoto (+3.62 pp). Overall, structural fairness depends on the interaction between graph degree structure and the representation interface.

Finding 4. Structural fairness is a deployment-relevant degree-position risk in graph prediction. In degree-skewed graphs, high-degree and low-degree nodes can experience different error profiles, so fairness assessment should report both the size and the direction of head–tail disparity.

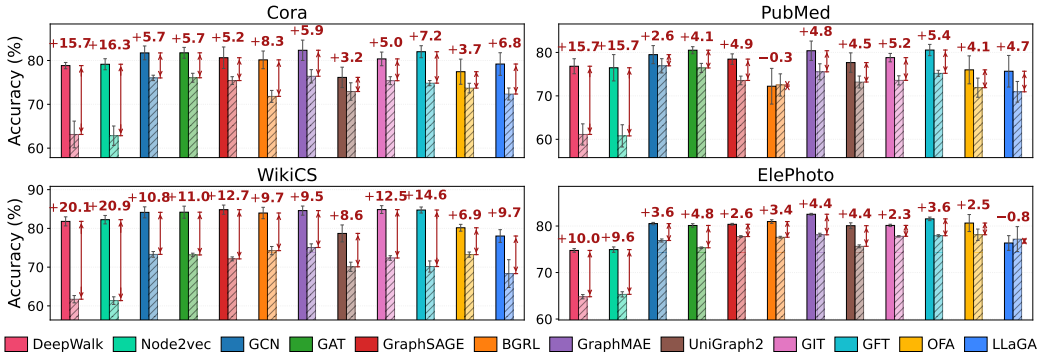


Figure 5: **Structural head–tail fairness.** Accuracy across all 12 methods on four node-classification datasets. Each method has two bars: head nodes (solid, top-degree quintile) and tail nodes (hatched, bottom-degree quintile). Red labels report the signed gap $\Delta = \mathrm{Acc}_{\mathrm{head}} - \mathrm{Acc}_{\mathrm{tail}}$ in percentage points; positive values indicate head-favoring behavior and negative values indicate tail-favoring behavior. Error bars show standard deviation across 5 seeds.

3.2.5 INTERPRETATION: ATTRIBUTION FIDELITY

Interpretation evaluates whether the input elements highlighted by a post-hoc attribution method are actually predictive. This axis does not ask whether a method is accurate, but whether its attributed substructures support the prediction better than a random baseline at the same sparsity.

Benchmark. We evaluate node-level attribution fidelity using gradient-saliency edge attribution under subgraph ablation (Agarwal et al., 2023). For each test node, we mask the top-k=10% edges in its K=2-hop receptive field and compute the GraphFramEx characterization score (Amara et al., 2022), the harmonic mean of Fid^{+} (confidence drop when salient edges are masked) and $1-\mathrm{Fid}^{-}$ (prediction preservation when non-salient edges are masked), computed on clipped fidelity terms (Appendix D.5). We report $\Delta_{\mathrm{char}} \times 100 = 100\,(\mathrm{char}_{\mathrm{sal}} - \mathrm{char}_{\mathrm{rand}})$, the lift over a random-edge baseline; a positive value means saliency identifies more predictive edges than guessing. This evaluation covers eight methods on eight NC datasets; DeepWalk and Node2vec lack an edge-conditioned differentiable prediction path, while OFA and LLaGA do not expose a comparable per-edge gradient interface under this protocol. Graph-level attribution results on molecular datasets are reported in Appendix G.5.

Evidence. Table 3 shows that node-level attribution fidelity is concentrated in a small set of methods. GFT has the highest cross-dataset mean (31.30), followed by GraphSAGE (27.23) and GIT (24.26); the remaining methods are at or below 14.00 on average. Per dataset, GFT leads six of eight columns, while GIT leads the two e-commerce graphs. Overall, node-level attribution fidelity is concentrated rather than universal: only a small subset of methods consistently produces edge saliency that outperforms random-edge ablation.

Finding 5. Explanation availability is not the same as attribution fidelity. For graph domains where explanations are used as scientific or molecular evidence, reliability evaluation should test whether highlighted edges or subgraphs are actually predictive under the stated attribution protocol, rather than assuming that a saliency interface yields faithful explanations.

3.3 CROSS-AXIS DISCUSSION

The three insights below are also supported on pretraining-unseen graphs (Appendix E).

Table 3: **Node-level subgraph-ablation fidelity.** $\Delta_{\text{char}} \times 100$ is the lift of gradient-saliency attribution over random-edge ablation at top-k=10%; higher is better. Cells report mean±std over runs, with the std of the lift propagated from the saliency and random scores; Avg reports the cross-dataset mean±std. **Bold**/underline: best/second per column.

Method	Cora	Citeseer	PubMed	WikiCS	Elec-Computers	Elec-Photo	Amazon-Ratings	Tolokers	Avg↑
GCN	$11.56_{\pm 2.45}$	$5.43_{\pm 6.35}$	$12.66_{\pm 2.85}$	$23.72_{\pm 2.24}$	$16.13_{\pm 6.05}$	$16.53_{\pm 4.92}$	$3.53_{\pm 3.57}$	$22.42_{\pm 0.29}$	$14.00_{\pm 7.24}$
GAT	$11.25_{\pm 1.97}$	$3.40_{\pm 1.95}$	$18.44_{\pm 3.63}$	$25.82_{\pm 4.57}$	$8.12_{\pm 5.41}$	$9.61_{\pm 4.36}$	$2.01_{\pm 4.04}$	$0.17_{\pm 0.03}$	$9.85_{\pm 8.71}$
GraphSAGE	$\underline{30.21}_{\pm 1.68}$	$11.97_{\pm 8.56}$	$\underline{33.74}_{\pm 4.13}$	$51.84_{\pm 1.65}$	$\underline{25.69}_{\pm 7.09}$	$\underline{24.83}_{\pm 6.58}$	$\underline{16.00}_{\pm 5.70}$	$\underline{23.56}_{\pm 0.61}$	$\underline{27.23}_{\pm 12.17}$
BGRL	$14.24_{\pm 1.75}$	$8.95_{\pm 0.75}$	$11.02_{\pm 4.82}$	$20.14_{\pm 2.34}$	$18.22_{\pm 7.38}$	$16.98_{\pm 4.26}$	$5.89_{\pm 5.57}$	$8.16_{\pm 1.77}$	$12.95_{\pm 5.21}$
GraphMAE	$12.81_{\pm 0.62}$	$6.58_{\pm 4.45}$	$17.34_{\pm 1.19}$	$25.36_{\pm 2.24}$	$9.13_{\pm 6.44}$	$10.55_{\pm 4.51}$	$6.71_{\pm 3.44}$	$1.21_{\pm 0.81}$	$11.21_{\pm 7.43}$
GFT	$\mathbf{30.64}_{\pm 3.48}$	$\mathbf{19.52}_{\pm 23.91}$	$\mathbf{40.11}_{\pm 5.19}$	$\mathbf{63.65}_{\pm 1.77}$	$24.87_{\pm 5.93}$	$24.17_{\pm 4.60}$	$\mathbf{17.25}_{\pm 2.31}$	$\mathbf{30.22}_{\pm 1.48}$	$\mathbf{31.30}_{\pm 14.90}$
GIT	$21.26_{\pm 18.59}$	$\underline{13.59}_{\pm 4.01}$	$30.12_{\pm 5.96}$	$\underline{53.50}_{\pm 3.07}$	$\mathbf{26.54}_{\pm 6.29}$	$\mathbf{25.00}_{\pm 4.75}$	$14.22_{\pm 3.06}$	$9.85_{\pm 0.99}$	$24.26_{\pm 13.77}$
UniGraph2	$5.38_{\pm 2.47}$	$2.75_{\pm 3.40}$	$10.73_{\pm 7.48}$	$24.58_{\pm 4.49}$	$7.33_{\pm 6.11}$	$6.40_{\pm 2.56}$	$5.02_{\pm 3.69}$	$14.19_{\pm 3.95}$	$9.55_{\pm 7.05}$

Insight 1: Reliability tests are not interchangeable, even on the same graph factor. Random edge deletion and saliency-guided edge ablation both alter graph structure but yield different method profiles. On four pretraining-unseen graphs (Table 8), GraphSAGE and GAT have the same mean edge-deletion drop (3.0 points), yet GraphSAGE has over three times the attribution lift (19.6 vs. 5.8). GFT also achieves the highest attribution lift despite a larger edge-deletion loss. Thus, similar structural interventions can expose distinct failure modes.

Insight 2: Reliability gains are axis-specific, even for foundation-era methods. Foundation-era methods lead in some regimes but not across the benchmark. On pretraining-unseen graphs, GFT has the strongest node-level attribution fidelity but degrades heavily under feature noise, where GraphMAE is most stable. UniGraph2 is competitive for minority-class recall on multiple imbalance datasets.

Insight 3: Some reliability failures are shared across all method families. Degree shift and rare-class coverage are two such failures. Under degree shift on pretraining-unseen graphs, the average gaps of fully covered feature-consuming methods, from supervised GNNs to foundation-era models, remain around 10–11 accuracy points and differ by less than two points. Class imbalance similarly produces persistent minority-class degradation.

4 Limitations and Future Work

GRL-Reliability does not cover all dimensions relevant to GRL reliability. The current benchmark focuses on five evaluation axes, leaving concerns such as calibration, adaptive adversarial robustness, and privacy outside its scope. Method coverage is also necessarily selective and does not include all architecture families or backbones, so the empirical findings should be interpreted with respect to the evaluated method set. Demographic fairness is currently evaluated on only one compatible dataset, limiting the generality of conclusions on this axis. Finally, the pretraining-unseen analysis removes direct evaluation-graph overlap but does not fully isolate the effects of pretraining composition.

5 Conclusion

GRL-Reliability shows that reliability in graph representation learning is multi-dimensional and cannot be inferred from clean performance or method family alone. Four practical lessons follow from our results (Appendix F): (i) clean accuracy should not be used as a proxy for reliability; (ii) deployment-relevant risks should be evaluated directly, since performance on one reliability axis generally does not predict another; (iii) single-axis leaders should be interpreted with caution, as strong performance on one axis can coexist with weak performance elsewhere; and (iv) when methods are not statistically separated on an axis, the benchmark does not support a fine-grained ranking. We hope GRL-Reliability serves as a reliability diagnostic and a testbed for methods targeting the remaining failure regimes.

Reproducibility Statement

Code, evaluation protocols, and scripts are available at the anonymous repository linked in the abstract. Dataset statistics and preprocessing are described in Appendix A, method descriptions in Appendix B, pretraining and adaptation hyperparameters in Appendix C, and the evaluation protocol of each axis in Appendix D. All reported numbers are aggregated from the released per-dataset outputs.

AI Use Statement

We used generative AI tools to edit the text and to assist with parts of the analysis scripts and the discussion of results; all analyses and conclusions were decided and verified by the authors, who take full responsibility for this work. Generative AI was not used to generate data, develop theoretical models, or formulate hypotheses; mathematical claims, proofs, and translation are not applicable.

REFERENCES

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023.

Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. GraphFramEx: Towards systematic evaluation of explainability methods for graph neural networks. In *Learning on Graphs Conference (LoG)*, 2022.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. LLaGA: Large language and graph assistant. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Zhikai Chen, Haitao Mao, Jingzhe Liu, Yu Song, Bingheng Li, Wei Jin, Bahare Fatemi, Anton Tsitsulin, Bryan Perozzi, Hui Liu, and Jiliang Tang. Text-space graph foundation models: Comprehensive benchmarks and new insights. *arXiv preprint arXiv:2406.10727*, 2024b.

Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 21(6):1011–1061, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

Yushun Dong, Zhenyu Lei, Zaiyi Zheng, Song Wang, Jing Ma, Alex Jing Huang, Chen Chen, and Jundong Li. PyGDebias: A python library for debiasing in graph learning. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1019–1022, 2024a.

Yushun Dong, Song Wang, Zhenyu Lei, Zaiyi Zheng, Jing Ma, Chen Chen, and Jundong Li. A benchmark for fairness-aware graph learning. *arXiv preprint arXiv:2407.12112*, 2024b.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Jiarui Feng, Hao Liu, Lecheng Kong, Mingfang Zhu, Yixin Chen, and Muhan Zhang. TAGLAS: An atlas of text-attributed graph datasets in the era of large graph and language models, 2024.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 855–864, 2016.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Xiaoguang Guo, Zehong Wang, Jiazheng Li, Shawn Spitzel, Qi Yang, Kaize Ding, Jundong Li, and Chuxu Zhang. Generalizing gnns with tokenized mixture of experts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. UniGraph2: Learning a unified embedding space to bind multimodal graphs. In *Proceedings of the ACM Web Conference (WWW)*, 2025.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. GraphMAE: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mingxuan Ju, Tong Zhao, Qianlong Wen, Wenhao Yu, Neil Shah, Yanfang Ye, and Chuxu Zhang. Multi-task self-supervised graph neural networks enable stronger task generalization. In *International Conference on Learning Representations (ICLR)*, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. DeepRobust: A platform for adversarial attacks and defenses. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Demonstration Track*, 2021.

Ziming Li, Youhuan Li, Xinhuan Chen, Lei Zou, Yang Li, Xiaofeng Yang, and Hongbo Jiang. Newsp: A new search process for continuous subgraph matching over dynamic graphs. In *IEEE International Conference on Data Engineering(ICDE)*, 2024.

Ziming Li, Youhuan Li, Yuyu Luo, Guoliang Li, and Chuxu Zhang. Graph neural networks for databases: A survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence(IJCAI)*, 2025.

Ziming Li, Xiaoming Wu, Zehong Wang, Jiazheng Li, Yijun Tian, Jinhe Bi, Yunpu Ma, Yanfang Ye, and Chuxu Zhang. Graph is a substrate across data modalities. In *International Conference on Machine Learning (ICML)*, 2026.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Xiaorui Liu, Jiayuan Ding, Wei Jin, Han Xu, Yao Ma, Zitao Liu, and Jiliang Tang. Graph neural networks with adaptive residual. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yihong Ma, Yijun Tian, Nuno Moniz, and Nitesh V. Chawla. Class-imbalanced learning on graphs: A survey. *ACM Computing Surveys*, 57(8), 2025. doi: 10.1145/3718734.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 701–710, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. doi: 10.1016/j.cell.2020.01.021.

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 *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2024.

Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Yiqi Wang, Jiliang Tang, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. Investigating and mitigating degree-related biases in graph convolutional networks. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1435–1444, 2020.

Komal K. Teru, Etienne Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning (ICML)*, 2020.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2022.

Yijun Tian, Kaiwen Dong, Chunhui Zhang, Chuxu Zhang, and Nitesh V Chawla. Heterogeneous graph masked autoencoders. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 37, pp. 9997–10005, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations (ICLR)*, 2019.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Song Wang, Zhen Tan, Yaochen Zhu, Chuxu Zhang, and Jundong Li. Generative risk minimization for out-of-distribution generalization on graphs. *Transactions on Machine Learning Research*, 2025a. arXiv:2502.07968.

Song Wang, Yushun Dong, Binchi Zhang, Zihan Chen, Xingbo Fu, Yinhan He, Cong Shen, Chuxu Zhang, Nitesh V Chawla, and Jundong Li. Safety in graph machine learning: Threats and safeguards. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Zehong Wang, Zheyuan Zhang, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. GFT: Graph foundation model with transferable tree vocabulary. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Zehong Wang, Zheyuan Liu, Tianyi Ma, Jiazheng Li, Zheyuan Zhang, Xingbo Fu, Yiyang Li, Zhengqing Yuan, Wei Song, Yijun Ma, Qingkai Zeng, Xiusi Chen, Jianan Zhao, Jundong Li, Meng Jiang, Pietro Lio, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116*, 2025b.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Towards graph foundation models: Learning generalities across graphs via task-trees. In *International Conference on Machine Learning (ICML)*, 2025c.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V Chawla, Chuxu Zhang, and Yanfang Ye. Beyond message passing: Neural graph pattern machine. In *International Conference on Machine Learning*, pp. 65496–65517. PMLR, 2025d.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. Generative graph pattern machine. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025e.

Zhonghao Wang, Danyu Sun, Sheng Zhou, Haobo Wang, Jiapei Fan, Longtao Huang, and Jiajun Bu. NoisyGL: A comprehensive benchmark for graph neural networks under label noise. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations (ICLR)*, 2022.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

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.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xingtong Yu, Shenghua Ye, Ruijuan Liang, Chang Zhou, Hong Cheng, Xinming Zhang, and Yuan Fang. Evaluating progress in graph foundation models: A comprehensive benchmark and new insights. *arXiv preprint arXiv:2603.10033*, 2026.

Xiangchi Yuan, Yijun Tian, Chunhui Zhang, Yanfang Ye, Nitesh V Chawla, and Chuxu Zhang. Graph cross supervised learning via generalized knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4083–4094, 2024a. doi: 10.1145/3637528.3671830.

Xiangchi Yuan, Chunhui Zhang, Yijun Tian, Yanfang Ye, and Chuxu Zhang. Mitigating emergent robustness degradation while scaling graph learning. In *International Conference on Learning Representations*, 2024b.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pp. 793–803, 2019.

Qinkai Zheng, Xu Zou, Yuxiao Dong, Yukuo Cen, Da Yin, Jiarong Xu, Yang Yang, and Jie Tang. Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

A DATASETS

Dataset statistics. We use twenty-five text-attributed graphs spanning seven domains, sourced from TSGFM (Chen et al., 2024b) and TAGLAS (Feng et al., 2024). Per-dataset statistics (domain, task, node and edge counts, class count, source) are reported in Table 4. We use the canonical train/validation/test split provided by each source repository, and repeat each method-dataset cell with five random seeds that re-randomize model initialization. Because all datasets are text-attributed, we use Sentence-BERT (Reimers & Gurevych, 2019) to encode each canonical text attribute (paper abstract, product description, item title, atom symbol, or relation phrase) into a 768-dimensional vector, yielding a single feature space shared across methods, consistent with the view of graphs as a unifying substrate across heterogeneous data modalities (Li et al., 2026).

Table 4: Datasets used in GRL-Reliability. **Task**: NC = node classification, LP = link prediction, GC = graph classification. For graph-classification datasets, #Nodes and #Edges are aggregated across all graphs in the dataset, and #Classes denotes the number of tasks in multi-task molecular benchmarks. †: included in the shared pretraining corpus. ‡: derived from arXiv; excluded from the pretraining-unseen analysis (Appendix E).

Dataset	Domain	Task	#Nodes	#Edges	#Classes	Source
Cora†	Academic	NC	2,708	10,556	7	(Chen et al., 2024b)
CiteSeer	Academic	NC	3,186	8,450	6	(Chen et al., 2024b)
PubMed†	Academic	NC	19,717	88,648	3	(Chen et al., 2024b)
arXiv†	Academic	NC	169,343	2,315,598	40	(Chen et al., 2024b)
arXiv23‡	Academic	NC	46,198	77,726	40	(Chen et al., 2024b)
arXivYear‡	Academic	NC	169,343	2,315,598	5	(Chen et al., 2024b)
DBLP	Academic	NC	14,376	431,326	4	(Chen et al., 2024b)
Elec-Computers	E-commerce	NC	87,229	1,256,548	10	(Chen et al., 2024b)
ElePhoto	E-commerce	NC	48,362	873,793	12	(Chen et al., 2024b)
SportsFit	E-commerce	NC	173,055	3,020,134	13	(Chen et al., 2024b)
AmazonRatings	E-commerce	NC	24,492	186,100	5	(Chen et al., 2024b)
BookChild	Book	NC	76,875	2,325,044	24	(Chen et al., 2024b)
BookHis	Book	NC	41,551	503,180	12	(Chen et al., 2024b)
WikiCS†	Web platform	NC	11,701	431,726	10	(Chen et al., 2024b)
Tolokers	Web platform	NC	11,758	519,000	2	(Feng et al., 2024)
WN18RR†	Knowledge graph	LP	40,943	93,003	11	(Feng et al., 2024)
FB15K237†	Knowledge graph	LP	14,541	310,116	237	(Feng et al., 2024)
ML1M	Recommendation	LP	9,923	2,000,418	5	(Feng et al., 2024)
BACE	Molecule	GC	51,577	111,536	1	(Feng et al., 2024)
BBBP	Molecule	GC	49,068	105,842	1	(Feng et al., 2024)
ChemHIV†	Molecule	GC	1,049,163	2,259,376	1	(Feng et al., 2024)
CYP450	Molecule	GC	414,367	895,886	5	(Feng et al., 2024)
MUV	Molecule	GC	2,255,846	4,892,252	17	(Feng et al., 2024)
Tox21	Molecule	GC	145,459	302,190	12	(Feng et al., 2024)
ToxCast	Molecule	GC	161,002	330,180	588	(Feng et al., 2024)

Note. Pretrained methods additionally use ChemBLPre and ChemPCBA as pretraining-only molecular corpora. These corpora are excluded from downstream evaluation and therefore not counted among the 25 benchmark datasets in Table 4.

Licenses and credit. All datasets are public academic graphs accessed through the TSGFM (Chen et al., 2024b) and TAGLAS (Feng et al., 2024) collections and are used in accordance with the licenses of their original sources (for example, the OGB datasets (Hu et al., 2020) under their MIT-style benchmark license, WikiCS under CC-BY-SA, and MoleculeNet datasets under their respective open data licenses). Method implementations (DeepWalk, Node2vec, GCN, GAT, GraphSAGE, BGRL, GraphMAE, GFT, GIT, UniGraph2, OFA, LLaGA) are cited at first mention and adapted from the

authors' public repositories under their respective open-source licenses (MIT, Apache 2.0, or BSD as released); no license terms are modified.

B METHODS

This appendix lists the twelve graph representation models evaluated in GRL-Reliability, grouped by method family (§3.1). Pretraining and downstream-adaptation hyperparameters for each method are documented separately in Appendix C (Implementation details).

B.1 SHALLOW GRAPH EMBEDDINGS

DeepWalk (Perozzi et al., 2014). A shallow node embedding method that learns representations by applying Skip-gram to uniform random walks over the graph.

Node2vec (Grover & Leskovec, 2016). A shallow node embedding method that extends DeepWalk with biased second-order random walks interpolating between breadth- and depth-first neighbourhood exploration.

B.2 SUPERVISED GNNS

GCN (Kipf & Welling, 2017). A message-passing GNN that aggregates neighbour features through symmetric normalised graph convolutions.

GAT (Veličković et al., 2018). A GNN that aggregates neighbour features through learned multi-head attention weights.

GraphSAGE (Hamilton et al., 2017). A GNN that generates node representations by sampling and aggregating features from a node's local neighbourhood.

B.3 SELF-SUPERVISED GRAPH MODELS

BGRL (Thakoor et al., 2022). A self-supervised graph representation method that learns by bootstrapping latent target predictions across two augmented views, without negative samples.

GraphMAE (Hou et al., 2022). A self-supervised graph representation method that uses masked-feature reconstruction with a scaled cosine error to pretrain a generative graph autoencoder.

B.4 GRAPH FOUNDATION MODELS

GFT (Wang et al., 2024a). A graph foundation model that converts cross-domain graph tasks into a shared computation-tree vocabulary and adapts to downstream tasks via prototype matching against learned tree codes.

GIT (Wang et al., 2025c). A graph foundation model pretrained with task-tree reconstruction over node-, edge-, and graph-level tasks, learning generalities across heterogeneous graphs.

UniGraph2 (He et al., 2025). A multi-domain graph foundation model that aligns features from different domains through routed mixture-of-experts on top of a shared embedding space.

OFA (Liu et al., 2024). A graph foundation model that unifies node, link, and graph tasks by reformulating each query as a binary link-prediction problem over an augmented prompt graph containing class and task-of-interest nodes.

LLaGA (Chen et al., 2024a). An LLM-based graph foundation model that projects a node's local neighbourhood into a Vicuna-7B token sequence and reads predictions from the language-model head.

C IMPLEMENTATION DETAILS

This appendix documents the pretraining and downstream-adaptation hyperparameter settings used in GRL-Reliability. Method descriptions are in Appendix B; per-axis evaluation protocols are in Appendix D.

C.1 PRETRAINING HYPERPARAMETERS

The two self-supervised methods (BGRL, GraphMAE) and the five graph foundation models (GFT, GIT, UniGraph2, OFA, LLaGA) pretrain on the same nine text-attributed graphs (cora, pubmed, arxiv, wikics, WN18RR, FB15K237, ChemHIV, ChemBLPre, ChemPCBA) adopted by GFT (Wang et al., 2024a) and OFA (Liu et al., 2024). Each method optimizes its native objective on this shared corpus, with hyperparameters following its source repository. The three supervised baselines (GCN, GAT, GraphSAGE) and the two topology-only methods (DeepWalk, Node2vec) have no pretraining stage; they are fitted directly on each downstream graph (see Appendix C.2). Per-method settings for the pretrained methods are as follows.

1. For BGRL (Thakoor et al., 2022): `num_layers=2, num_hidden=768, lr=1e-4, weight_decay=1e-5, max_epoch=50, batch_size=1024, ema_tau=0.99, warmup_frac=0.1, drop_edge=0.2, drop_feat=0.2, predictor_hidden=512, optimizer=AdamW`.
2. For GraphMAE (Hou et al., 2022): `num_layers=2, num_hidden=768, num_heads=4, lr=1e-3, weight_decay=2e-4, max_epoch=50, batch_size=1024, mask_rate=0.5, replace_rate=0.0, loss=sce, alpha_l=3.0, in_drop=0.2, attn_drop=0.1, optimizer=AdamW`.
3. For GFT (Wang et al., 2024a): `num_layers=2, hidden_dim=768, dropout=0.15, pretrain_lr=1e-4, pretrain_weight_decay=1e-5, pretrain_epochs=50, pretrain_batch_size=1024, codebook_size=128, codebook_head=4, codebook_decay=0.8, feat_p=0.2, edge_p=0.2, topo_recon_ratio=0.1, optimizer=AdamW`.
4. For GIT (Wang et al., 2025c): `num_layers=2, hidden_dim=768, dropout=0.15, lr=1e-7, weight_decay=1e-8, epochs=20, batch_size=4096, fanout=10, ema=0.99, feat_p=0.2, edge_p=0.2, topo_recon_ratio=0.1, optimizer=AdamW`.
5. For UniGraph2 (He et al., 2025): `num_layers=3, num_hidden=768, num_experts=8, num_selected_experts=2, lr=1e-4, weight_decay=1e-5, max_epoch=50, batch_size=64, gamma=2.0, lambda_spd=0.5, feat_drop_rate=0.1, edge_mask_rate=0.1, fanout=10, optimizer=AdamW`.
6. For OFA (Liu et al., 2024): `num_layers=7, num_hidden=768, dropout=0.15, lr=1e-3, num_epochs=50, batch_size=128, num_relations=5, JK=none, optimizer=AdamW`.
7. For LLaGA (Chen et al., 2024a), following the original two-stage recipe, only the projector is trained with the Vicuna-7B language model frozen: `llm_backbone=Vicuna-7B (frozen), projector_type=2-layer-mlp, projector_hidden=4096, lr=5e-4, weight_decay=1e-4, epochs=10, batch_size=4, sample_size=10, use_hop=2, max_text_len=128, max_steps=5000, dtype=bfloat16, optimizer=AdamW`.

C.2 DOWNSTREAM ADAPTATION HYPERPARAMETERS

For each method-dataset cell we apply the method's native downstream adaptation procedure where available, with five random seeds. Per-method settings are as follows.

1. For DeepWalk (Perozzi et al., 2014), walk embeddings are fitted on the downstream graph with Skip-gram on uniform random walks (p=q=1): `emb_dim=128,`

`walk_length=20, walks_per_node=10, context_size=10, num_negative=1, walk_epochs=100, batch_size=128, lr=1e-2, optimizer=SparseAdam`. Predictions use a frozen-feature ℓ_2-regularised logistic-regression head (scikit-learn, `C=1.0, max_iter=1000, solver=lbfgs`).

2. For Node2vec (Grover & Leskovec, 2016): same Skip-gram and logistic-regression configuration as DeepWalk, with biased second-order walks (`p=1.0, q=0.5`, BFS-leaning).
3. For GCN, GAT, and GraphSAGE: `num_layers=2, num_hidden=768, lr=1e-3, dropout=0.2, weight_decay=1e-4, max_epochs=500, patience=200`.
4. For BGRL (Thakoor et al., 2022): `num_layers=2, num_hidden=768, lr=5e-4, dropout=0.2, weight_decay=1e-5, max_epochs=500, patience=200`.
5. For GraphMAE (Hou et al., 2022): `num_layers=2, num_hidden=768, lr=1e-3, weight_decay=1e-4, dropout=0.2, max_epochs=500, patience=200`.
6. For UniGraph2 (He et al., 2025): `num_layers=3, num_hidden=768, num_experts=8, num_selected_experts=2, lr=1e-3, weight_decay=1e-4, dropout=0.2, max_epochs=500, patience=200`.
7. For GFT (Wang et al., 2024a): `num_layers=2, hidden_dim=768, dropout=0.15, finetune_lr=5e-4, finetune_epochs=1000, early_stop=200, batch_size=full`.
8. For GIT (Wang et al., 2025c): `num_layers=2, hidden_dim=768, dropout=0.15, lr=1e-4, sft_epochs=100, weight_decay=0, batch_size=full`.
9. For OFA (Liu et al., 2024): `num_layers=7, num_hidden=768, dropout=0.15, lr=1e-3, weight_decay=0, epochs=100, patience=30, prompt_graph_mode=subgraph, subgraph_hops=2`.
10. For LLaGA (Chen et al., 2024a): `lr=1e-3, weight_decay=1e-4, max_epochs=15, patience=3, sample_size=10, max_text_len=128, batch_size=8, dtype=bfloat16`.

For node classification, all methods use cross-entropy loss with their native prediction head: GFM methods (GFT/GIT/UniGraph2) use prototype or code-based classification; OFA uses class-node link scoring with binary cross-entropy; LLaGA uses the language-model head over a verbalised class vocabulary; the remaining methods use a linear classification head. For link prediction, methods without a native link-level head score node pairs with a dot-product decoder trained with binary cross-entropy on sampled negative edges. For graph classification, methods without a native graph-level head mean-pool node representations and use binary cross-entropy per task; on multi-task molecular datasets, missing labels are masked from the loss and AUC is averaged over tasks. For inductive KG completion, all evaluated methods share a DistMult decoder trained with head and tail corruption negatives under a logistic loss, and MRR and Hits@10 are computed with filtered ranking, averaged over head and tail prediction. No class reweighting, focal loss, or oversampling is applied.

C.3 Compute resources

All experiments were run on a shared university HPC cluster (SLURM scheduler, `general-gpu` partition) with single-GPU jobs. The GPU pool consists of NVIDIA L40 (48 GB VRAM) and NVIDIA A100 (40 GB VRAM) accelerators, allocated with 4–8 CPU cores and 32–192 GB of system memory per job depending on dataset size.

Per method-dataset-seed cell, wall-clock time ranges from approximately 5–30 minutes on small node-classification graphs (Cora, CiteSeer, PubMed, ArXiv23, DBLP, AmazonRatings, WikiCS, Tolokers, and BookHis), 30 minutes to 2 hours on medium graphs and link-prediction datasets (ElePhoto, Elec-Computers, BookChild, WN18RR, FB15K237, and ML1M) and on smaller molecular graph-classification corpora (BACE, BBBP, Tox21, and ToxCast), and 2–6 hours on large node-classification

graphs (arXiv, arXivYear, and SportsFit) and larger molecular graph-classification corpora (CYP450, ChemHIV, and MUV). Dataset statistics are reported in Table 4. Joint-corpus pretraining of each SSL backbone is a one-off cost on the order of 8–24 GPU-hours.

The total compute for the experiments reported in this paper is approximately on the order of 10^4 GPU-hours, aggregating across methods, datasets, evaluation axes, and seeds. The full research project, including preliminary configurations and discarded intermediate runs, required additional compute beyond the final reported numbers.

D Evaluation axis details

This appendix gives, for each of the five evaluation axes of GRL-Reliability, the per-axis evaluation protocol referenced in §3.2. Method-side hyperparameters are in Appendix C; per-axis extended numerical results are in Appendix G. Table 5 summarizes the five evaluation protocols. Structurally related operations can measure different properties: random edge deletion measures prediction stability, whereas saliency-guided edge ablation measures evidence fidelity.

Table 5: Summary of the five evaluation protocols.

Axis	Setup	Measures	Metric	Direction
Corruption	Feature noise; edge deletion	Stability	Performance drop	↓
OOD	Degree/temporal/scaffold/entity shift	Generalization	ID–OOD gap / MRR	↓ / ↑
Imbalance	Minor-class downsampling	Rare-class coverage	Minor-class recall	↑
Fairness	Structural/demographic groups	Group disparity	|Group gap|	↓
Interpretation	Saliency vs. random ablation	Evidence fidelity	$\Delta_{\text{char}} \times 100$	↑

Dataset coverage is axis-specific. Degree-shift OOD is evaluated on 13 node-classification datasets, with the main text showing six representative datasets and the appendix reporting all 13. Class imbalance is evaluated on eight categorical-label node-classification datasets, while structural fairness is evaluated on eight node-classification datasets with degree-based head-tail groups. Interpretation is evaluated on eight node-classification datasets for methods whose adaptation interface exposes edge-level attribution. Temporal, scaffold, and inductive-entity shifts are defined on datasets that provide the corresponding time, molecular scaffold, or entity split.

Table 6 summarises which (method, axis) cells are evaluated. All twelve methods are evaluated on the corruption, OOD, imbalance, and fairness axes, with sub-condition restrictions for the topology-only methods: DeepWalk and Node2vec consume only graph structure, so under corruption they are evaluated on the structure-noise channel only (feature noise is inapplicable), and under OOD they are evaluated on the node-classification shifts (degree and temporal) but not on the graph-classification scaffold shift or the link-prediction inductive-entity shift, which require task-format adaptations they do not provide. The interpretation axis is restricted to the eight feature-consuming non-LLM methods: OFA's prompt-graph reformulation reframes each query as binary link prediction over an augmented prompt graph, which is not directly compatible with the edge-saliency operator on the original graph; LLaGA's current adaptation interface does not expose a validated gradient path that can be aligned with edges of the original graph; and DeepWalk/Node2vec's inference path does not condition on individual edges (random walks fold the graph into static node embeddings before downstream prediction).

D.1 Corruption

We evaluate robustness to two complementary input-corruption operationalizations applied to the evaluation graph: *feature noise*, which perturbs node attributes (Liu et al., 2021), and *structure noise*, which perturbs graph topology (Rong et al., 2020). All methods except DeepWalk and Node2vec are trained on the clean graph, and their parameters are frozen during corruption evaluation. For DeepWalk and Node2vec, embeddings and the downstream logistic-regression classifier are refitted on each edge-deleted graph, keeping the training/test splits and hyperparameters unchanged. Each method is evaluated under six severity levels (severity 0 = clean and five non-zero severities) with a deterministic per-(split, severity) random seed for reproducibility. Performance is measured with

Table 6: Coverage of (method, evaluation axis) cells in GRL-Reliability. ✓ indicates evaluation on at least one dataset within the corresponding axis; − = not evaluated (adaptation interface incompatible with the per-axis evaluation protocol; see paragraph above).

Method	Corruption	OOD	Imbalance	Fairness	Interpretation
DeepWalk	✓	✓	✓	✓	−
Node2vec	✓	✓	✓	✓	−
GCN	✓	✓	✓	✓	✓
GAT	✓	✓	✓	✓	✓
GraphSAGE	✓	✓	✓	✓	✓
BGRL	✓	✓	✓	✓	✓
GraphMAE	✓	✓	✓	✓	✓
GFT	✓	✓	✓	✓	✓
GIT	✓	✓	✓	✓	✓
UniGraph2	✓	✓	✓	✓	✓
OFA	✓	✓	✓	✓	−
LLaGA	✓	✓	✓	✓	−

classification accuracy on node-classification datasets, AUC-ROC on link-prediction datasets, and AUC on graph-classification datasets.

Feature noise. At inference time, independent zero-mean Gaussian noise is added to every node's feature vector, with the noise scaled per feature dimension by the training-set per-dimension standard deviation:

$$x_{\text{noisy}} = x + \sigma_{\text{rel}} \cdot \text{std}(X_{\text{train}}) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I).$$

The relative noise scale takes five values $\sigma_{\text{rel}} \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$. Edges, labels, train/validation/test splits, and the trained model parameters are not modified. Per-dataset severity curves across all ten feature-consuming methods are reported in Figure 6 (DeepWalk and Node2vec do not consume node features and are therefore excluded from feature-noise evaluation).

Structure noise. At inference time, a fraction p of undirected edges is removed uniformly at random; both directions of each selected edge are dropped jointly so that the perturbed graph remains symmetric, and self-loops (when present) are preserved. The drop rate takes five values $p \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$. Node features, labels, and train/validation/test splits are not modified. Per-dataset severity curves across all twelve methods are reported in Figure 7.

D.2 Out-of-distribution generalization

We evaluate four complementary graph-context shifts that target different aspects of the input distribution: *degree shift* on node attributes' structural neighbourhood, *temporal shift* on publication time, *scaffold shift* on molecular substructure, and *inductive KG shift* on the entity vocabulary visible at test time. Following the spirit of mechanism-level evaluation in graph OOD benchmarks (Gui et al., 2022), in every case the model is trained on the in-distribution (ID) training partition and evaluated on a held-out out-of-distribution (OOD) partition that shares the task definition but differs in the targeted distribution.

Degree shift. For each of the 13 node-classification datasets in Table 12, labelled nodes are sorted by node degree in descending order; the top 60% form the training pool, the next 20% form the OOD validation set, and the bottom 20% form the OOD test set. Within the training pool, randomly held-out subsets of 10% of the labelled nodes serve as the ID validation and ID test sets, and acc_{ID} is measured on the latter. The protocol matches GOOD's covariate-shift setup (Gui et al., 2022). We report the ID-to-OOD accuracy drop $\Delta = \text{acc}_{\text{ID}} - \text{acc}_{\text{OOD}}$ per (method, dataset) cell. Per-dataset heatmaps for the representative six-dataset subset are reported in Figure 3a (main text), with full per-dataset results in Appendix G.2.

Temporal shift. On the arXiv citation network, training uses papers with publication year ≤ 2010 and OOD evaluation uses papers from ≥ 2017, with intermediate years 2011–2016 reserved as OOD validation. This split builds on the arXiv temporal split of Wu et al. (2022) and reflects the kind of evolving-graph context studied in continuous subgraph search over dynamic graphs (Li et al., 2024) and, more broadly, in graph neural networks for database systems (Li et al., 2025). We report the ID-to-OOD accuracy drop sorted across the twelve methods in Figure 3b.

Scaffold shift. On the molecular graph-classification datasets BACE and Tox21, we compare random splits against Bemis–Murcko scaffold splits (Hu et al., 2020): molecules are grouped by their Bemis–Murcko scaffold (computed from the canonical SMILES) and the train, validation, and test sets receive disjoint scaffold groups in an $80:10:10$ ratio. We report the AUC drop $\Delta\mathrm{AUC} = \mathrm{AUC}_{\text{rand}} - \mathrm{AUC}_{\text{scaffold}}$, so a smaller $\Delta\mathrm{AUC}$ indicates better generalization to unseen scaffolds.

Inductive KG shift. On the knowledge-graph link-prediction datasets WN18RR and FB15K237, we follow the fully inductive setting of Teru et al. (2020). The entity set is randomly partitioned into disjoint training and test pools ($75:25$). Facts with both endpoints in the training pool form the training graph, facts with both endpoints in the test pool form the test graph, and facts crossing the two pools are discarded. Test-graph facts are further split $80:10:10$ into a support graph, validation queries, and test queries. At inference, test entities are encoded from their Sentence-BERT text features and the support graph, which is the only structure visible for test entities; query facts are never used for message passing. For each test query (h, r, t), the true tail is ranked against all test-pool entities, and symmetrically for the head, with other true triples of the test graph removed from the candidate list (filtered setting). Training negatives corrupt heads or tails with random training-pool entities. We report MRR (mean reciprocal rank of the correct entity) and Hits@10 (fraction of queries whose correct entity is ranked in the top 10), averaged over head and tail prediction. Test entities are unseen during downstream training and require the method to generalize beyond the entity vocabulary observed at training time.

D.3 Class imbalance

We follow a step-imbalance protocol on each node-classification dataset. Classes are sorted in ascending order of their original training-set sample count; the lower half are designated as *minor* classes and the upper half as *major* classes. Let $n_{\text{major}} = \max_{c \in \text{major}} n_c$ denote the maximum training count among major classes. Each minor class is then randomly downsampled to

$$n_{\text{minor}} = \max\bigl(1,\ \lfloor n_{\text{major}}/\rho \rfloor\bigr)$$

training instances at imbalance ratio $\rho \in \{5, 10, 20\}$, with the major-class training samples, the validation set, and the test set left untouched. The per-cell sampling is deterministic given the seed. We report per-class recall on the test split, averaged separately across major and minor classes. Per-dataset per-class recall at all three imbalance ratios is reported in Appendix G.3.

D.4 Fairness

We evaluate two complementary fairness operationalizations: *structural fairness*, which measures the disparity in prediction quality between high- and low-degree nodes, and *demographic fairness*, which measures the disparity between groups defined by a protected demographic attribute. In both cases the model is trained without any fairness-aware modification (no group reweighting, no constrained loss), so the measurements reflect each method's group disparity under our unified fine-tuning protocol.

Structural fairness. Following the head/tail degree-disparity formulation of Tang et al. (2020), on each of the eight node-classification datasets in Appendix G.4, test nodes are partitioned by degree quantile: with $q = 0.2$, the highest-degree 20% of test nodes form the *head* group, and the lowest-degree 20% form the *tail* group; the middle 60% is not assigned to either group for the gap measurement. We report per-group test accuracy ($\mathrm{acc}_{\text{head}}$, $\mathrm{acc}_{\text{tail}}$) and the head$-$tail accuracy gap $\Delta_{\text{HT}} = \mathrm{acc}_{\text{head}} - \mathrm{acc}_{\text{tail}}$. A larger positive gap indicates a more pronounced low-degree disadvantage. Per-dataset results across the eight node-classification datasets are reported in Appendix G.4.

Demographic fairness. Demographic fairness is evaluated on the Tolokers banned-user-prediction dataset, with the user education attribute as a binary sensitive attribute (s=1 for users with higher education, s=0 otherwise) following the protocol of Dong et al. (2024a). We report three group-disparity quantities at test time:

$$\begin{aligned}\Delta\mathrm{SP} &= |P(\hat{y}{=}1 \mid s{=}0) - P(\hat{y}{=}1 \mid s{=}1)|,\\ \Delta\mathrm{EO} &= |P(\hat{y}{=}1 \mid s{=}0,\, y{=}1) - P(\hat{y}{=}1 \mid s{=}1,\, y{=}1)|,\\ \Delta\mathrm{Util} &= |\mathrm{AUC}(s{=}0) - \mathrm{AUC}(s{=}1)|.\end{aligned}$$

ΔSP measures statistical-parity disparity, ΔEO measures equal-opportunity disparity, and $\Delta\mathrm{Util}$ measures the per-group AUC gap. Smaller values indicate smaller group disparity. Per-method results are reported in Appendix G.4.

D.5 Interpretation

We evaluate post-hoc explanation fidelity at two task levels: *node-level interpretation* on node-classification datasets via edge-saliency subgraph ablation, and *graph-level interpretation* on molecular graph-classification datasets via atom-saliency node ablation. Both follow the GraphFramEx characterization-score protocol of Amara et al. (2022).

Node-level interpretation. For each test node we restrict the receptive field to the K-hop computational subgraph (K=2) and compute per-edge saliency by aggregating the gradient-based input attribution at its two endpoints:

$$S_{(u,v)} = \left\|\partial \operatorname{logit}_{\hat{y}}/\partial x_u\right\| + \left\|\partial \operatorname{logit}_{\hat{y}}/\partial x_v\right\|.$$

Edges are then ranked by $S_{(u,v)}$ and we evaluate at four sparsity levels $k \in \{5, 10, 20, 50\}\%$. With p_0 the clean predicted-class probability and p_{+k}, p_{-k} the probabilities after masking the top-k and the complementary edges respectively, we report

$$\begin{aligned}\mathrm{Fid}^{+} &= p_0 - p_{+k} \quad \text{(higher better)},\\ \mathrm{Fid}^{-} &= p_0 - p_{-k} \quad \text{(lower better)},\\ \mathrm{char} &= \frac{2 \cdot \mathrm{Fid}^{+} \cdot (1 - \mathrm{Fid}^{-})}{\mathrm{Fid}^{+} + (1 - \mathrm{Fid}^{-}) + \epsilon} \quad \text{(higher better)},\end{aligned}$$

where char is the GraphFramEx characterization score, a harmonic-style combination of Fid^{+} and $(1 - \mathrm{Fid}^{-})$ with values in $[0, 1]$. In our implementation, per-node fidelity values are clipped to $[0, 1]$ before computing the characterization score, which is then averaged across nodes; the reported Fid^{+} and Fid^{-} values are unclipped. To control for saliency-method specificity, every metric is reported for both gradient saliency and a uniform-random baseline, and the headline *lift* is $\Delta\mathrm{char} = \mathrm{char}_{\text{grad}} - \mathrm{char}_{\text{random}}$. The full per-dataset sparsity sweep and the per-dataset $\mathrm{Fid}^{+}/\mathrm{Fid}^{-}$ decomposition at k=10% are reported in Appendix G.5.

Graph-level interpretation. For each test molecule we compute per-atom saliency as the L_1-norm gradient of the predicted-class logit with respect to that atom's input features. Atoms are then ranked by saliency and the same Fidelity quantities are computed by ablating the top-k or the complementary atoms (atom removal also drops the incident bonds and pool entries). Edge saliency, when needed for the subgraph ablation, is derived as the sum of the two endpoint atom saliencies. Per-method graph-level interpretation results on the molecular datasets are reported in Appendix G.5.

E Analysis on Pretraining-Unseen Graphs

Setting. Because several evaluation graphs also appear in the shared pretraining corpus, we additionally report results on pretraining-unseen graphs. Seven of the twenty-five evaluation datasets belong to this corpus (marked in Table 4); we conservatively also exclude the two arXiv-derived datasets (arXiv23 and arXivYear), leaving sixteen pretraining-unseen datasets. Each axis is evaluated on its applicable unseen subset, as specified in the table captions, using the same runs as the full benchmark.

Table 7: Corruption and interpretation on pretraining-unseen graphs. FN / ED: drop from clean to the highest feature-noise / edge-deletion severity, averaged over the 13 unseen datasets evaluated for all eight methods (CiteSeer, DBLP, ElecComp, ElePhoto, SportsFit, AmazonRatings, BookChild, BookHis, BACE, CYP450, MUV, Tox21, ToxCast; accuracy for node classification, ROC-AUC for graph classification; lower is better). Lift: attribution lift $\Delta_{\text{char}} \times 100$ over the random-edge baseline on the 5 unseen interpretation datasets (CiteSeer, ElecComp, ElePhoto, AmazonRatings, Tolokers; higher is better).

Method	**FN drop** $\downarrow$	**ED drop** $\downarrow$	**Lift** $\uparrow$
GCN	7.9	3.3	12.8
GAT	11.8	3.5	4.7
GraphSAGE	9.0	4.9	20.4
BGRL	9.7	9.5	11.6
GraphMAE	3.8	4.7	6.8
UniGraph2	8.1	3.7	7.1
GIT	16.3	6.3	17.8
GFT	13.6	9.4	23.2

Table 8: Corruption and interpretation on the four pretraining-unseen graphs covered by both axes (CiteSeer, ElecComp, ElePhoto, AmazonRatings). Metrics as in Table 7.

Method	**FN drop** $\downarrow$	**ED drop** $\downarrow$	**Lift** $\uparrow$
GCN	10.9	3.0	10.4
GAT	15.4	3.0	5.8
GraphSAGE	11.2	3.0	19.6
BGRL	10.1	3.5	12.5
GraphMAE	5.9	3.2	8.2
UniGraph2	11.3	3.5	5.4
GIT	18.5	4.3	19.8
GFT	13.2	4.9	21.5

Table 9: Degree shift, class imbalance, and structural fairness on pretraining-unseen graphs. Degree: mean $|\text{ID} - \text{OOD}|$ accuracy gap on 8 unseen datasets (CiteSeer, DBLP, ElecComp, ElePhoto, SportsFit, AmazonRatings, BookChild, BookHis). Minor-recall drop: decrease in minor-class recall from ρ=5 to ρ=20 on 5 unseen datasets (CiteSeer, DBLP, ElecComp, ElePhoto, SportsFit). Head–tail gap: mean $|\text{Acc}_{\text{head}} - \text{Acc}_{\text{tail}}|$ on 5 unseen datasets (CiteSeer, ElecComp, ElePhoto, AmazonRatings, Tolokers). Lower indicates a smaller gap or degradation; minor-recall drops should be read together with the starting recall because of floor effects. *Covers 4 of the 8 (degree) or 4 of the 5 (imbalance) datasets.

Method	**Degree gap** $\downarrow$	**Minor-recall drop** $\downarrow$	**Head–tail gap** $\downarrow$
DeepWalk	18.1	7.1	14.9
Node2vec	18.2	6.8	14.8
GCN	10.0	10.7	10.3
GAT	11.3	12.0	11.4
GraphSAGE	10.8	8.5	11.0
BGRL	9.9	11.0	11.4
GraphMAE	9.8	11.3	12.2
UniGraph2	10.8	11.1	11.2
GIT	11.0	9.4	10.9
GFT	11.0	11.6	12.9
OFA	10.4	22.4*	5.2
LLaGA	8.0*	12.8	8.8

Results. The unseen-graph analysis shows qualitatively similar behavior to the full benchmark. Corruption robustness remains method-specific rather than family-wide: GraphMAE has the smallest feature-noise drop, whereas GIT and GFT have the largest, and BGRL and GFT degrade most under edge deletion. GFT keeps the highest attribution lift, exceeding GCN, GAT, BGRL, GraphMAE, and UniGraph2 on all five unseen interpretation datasets. Under degree shift, topology-only embeddings still show the largest gap, about 1.6–$1.9\times$ that of the nine fully covered feature-consuming methods, whose average gaps lie within 1.5 points of each other. Imbalance and head–tail disparities remain method- and dataset-dependent.

Statistical support. We use exact two-sided Wilcoxon signed-rank tests on per-dataset seed means, with Holm correction within each comparison family. Table 10 reports the headline corruption comparisons on the 13 pretraining-unseen graphs of Table 7; all remain significant after correction. Friedman tests also reject equal performance across the eight methods for feature noise (χ^2=34.0, $p<10^{-4}$) and edge deletion (χ^2=18.8, p=0.009). For degree shift, none of the nine topology-only versus feature-consuming comparisons remains significant after Holm correction (minimum adjusted p=0.070), so we treat this pattern as descriptive.

Table 10: Statistical support for headline corruption comparisons on pretraining-unseen graphs. Mean difference: first method minus second method.

Condition	**Comparison**	**Mean difference**	p	**Adjusted** p
Feature noise	GFT vs. GCN	$+5.7$	0.006	0.019
Feature noise	GIT vs. GCN	$+8.4$	0.017	0.034
Feature noise	GFT vs. GraphMAE	$+9.9$	0.0002	0.002
Feature noise	GraphMAE vs. GCN	-4.2	0.021	0.034
Edge deletion	GFT vs. GCN	$+6.1$	0.002	0.012
Edge deletion	BGRL vs. GCN	$+6.2$	0.005	0.019

F Cross-Axis Reliability Analysis

This appendix supports the practical lessons in the conclusion. We restrict the analysis to the eight methods evaluated on every axis (DeepWalk and Node2vec have no feature-noise or attribution results, and OFA and LLaGA have no attribution results) and, for each axis or pair of axes, to the datasets on which all eight methods have results. All quantities are computed from the results reported in the main text and in Appendices E and G. Metrics are oriented so that higher is better (negated drops and gaps).

Table 11: Average rank (1 = best) of the eight methods evaluated on every axis, computed per dataset and averaged over each axis's datasets. The last rows give the number of datasets and the Friedman test p-value for differences among the eight methods.

Method	**Clean**	**Feat. noise**	**Edge del.**	**Degree**	**Imbalance**	**Fairness**	**Attribution**
GCN	4.1	3.1	**3.2**	4.0	5.2	**3.5**	5.6
GAT	4.9	4.9	3.4	6.1	4.0	4.8	6.5
GraphSAGE	3.7	4.6	4.9	4.8	5.0	4.4	2.2
BGRL	3.6	3.8	4.5	**2.9**	4.9	4.5	5.2
GraphMAE	**3.3**	**2.2**	3.6	3.4	4.8	4.8	5.4
UniGraph2	7.3	4.5	5.3	5.2	**2.9**	**3.5**	7.0
GIT	3.7	6.2	4.6	4.5	5.8	4.0	2.5
GFT	5.5	6.7	6.4	5.1	3.5	6.6	**1.5**
Datasets	19	19	20	13	8	8	8
Friedman p	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-3}$	0.028	0.28	0.26	$< 10^{-4}$

Rank agreement. For each pair of metrics, we compute Kendall's τ between the rankings of the eight methods within each shared dataset and average over datasets, with 95% bootstrap intervals

over datasets. Clean accuracy agrees little with every reliability axis (mean τ from -0.01 to 0.13, all intervals including zero). Among reliability axes, only the two corruption channels agree moderately ($\tau = 0.35$, $[0.22, 0.48]$, 19 datasets), and edge deletion and degree shift agree weakly ($\tau = 0.21$, $[0.01, 0.40]$, 13 datasets); all other pairs have $|\tau| \leq 0.25$ with intervals including zero.

Interpretation. Clean-accuracy rankings thus carry little information about reliability rankings, and reliability axes largely disagree with one another. Axis leaders can rank poorly elsewhere: GFT ranks first on attribution but 6.7 and 6.4 on feature noise and edge deletion, and GraphMAE ranks first on feature noise but 5.4 on attribution (Table 11). Finally, clean accuracy, both corruption channels, and attribution separate the methods strongly (Friedman $p < 10^{-3}$), whereas degree shift separates them only weakly ($p = 0.028$) and imbalance and structural fairness do not ($p = 0.28$ and 0.26); on these axes the current evidence does not support a fine-grained ranking.

G Extended Experimental Results

G.1 Corruption

G.1.1 FEATURE NOISE

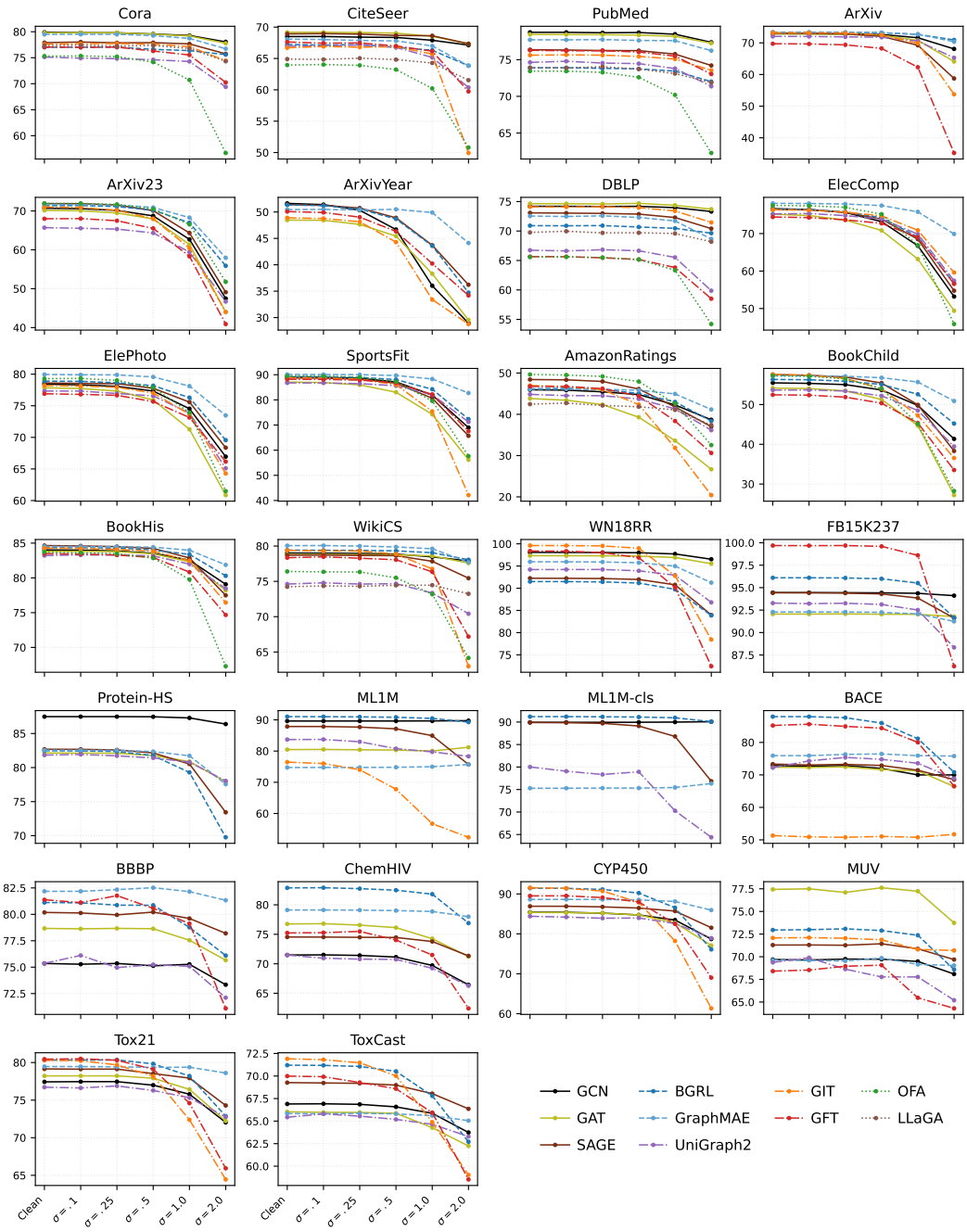


Figure 6: Per-dataset feature-noise severity curves on GRL-Reliability datasets. Each panel reports test accuracy (NC) or AUC (LP, GC) at clean and the five relative noise scales ($\sigma_{\text{rel}} \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$).

G.1.2 EDGE DELETION

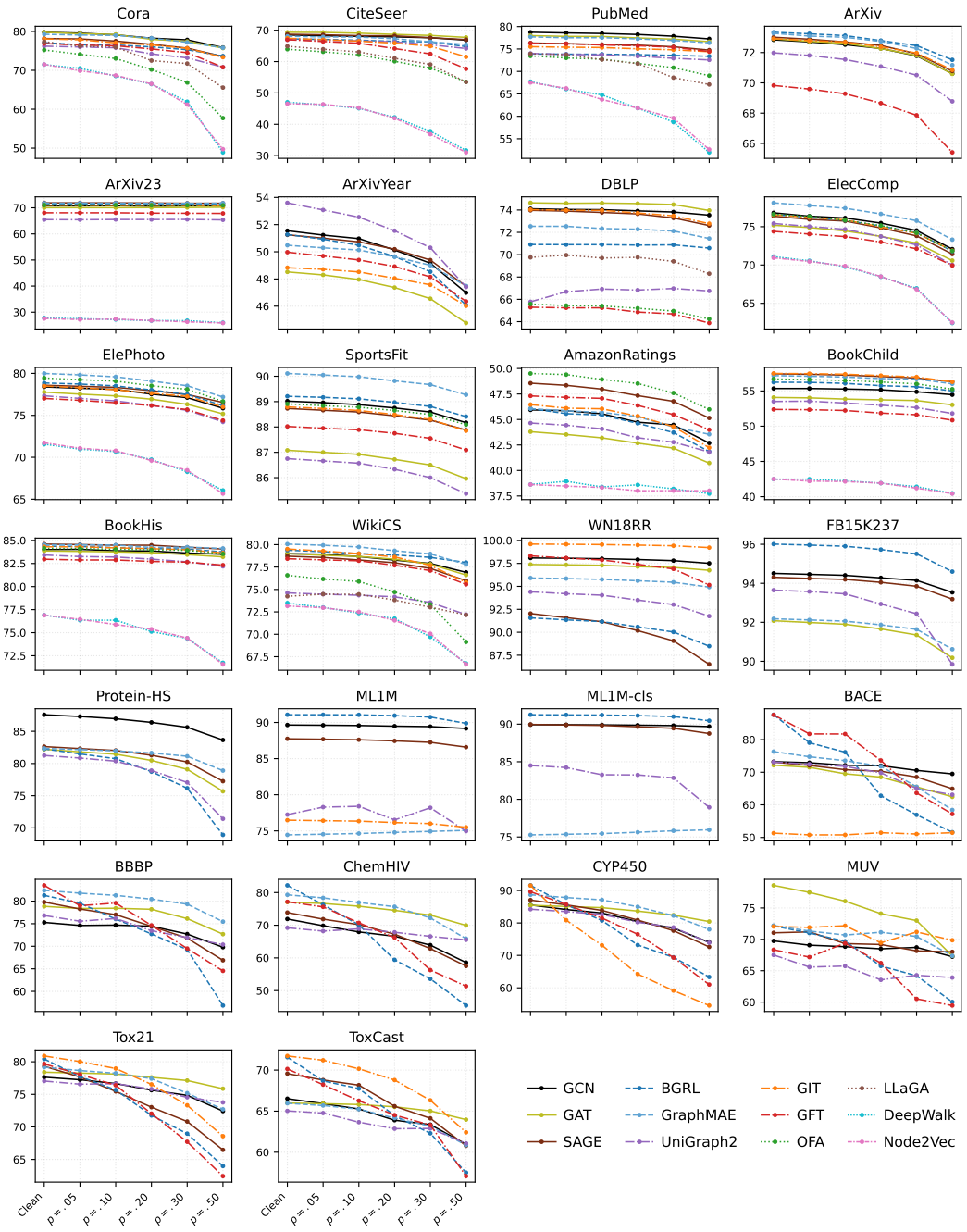


Figure 7: Per-dataset edge-deletion severity curves on GRL-Reliability datasets. Each panel reports test accuracy (NC) or AUC (LP, GC) at clean and the five edge-drop probabilities ($p \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$).

G.2 Out-of-Distribution Generalization — full results

Table 12: Full per-dataset OOD degree-shift results across all 13 datasets (mean±std over 5 seeds). Each block shows up to four datasets side by side; per (dataset, column), **bold** = best; underline = second. Gap = ID − OOD, smaller |Gap| more robust. "–" indicates unavailable results for the corresponding method–dataset setting.

Method	cora			citeseer			pubmed			arxiv		
	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap
DeepWalk	86.79±1.57	66.49±1.04	+20.30	71.22±1.12	40.75±1.98	+30.47	83.13±0.36	73.14±0.55	+9.99	73.67±0.24	49.14±0.16	+24.53
Node2Vec	87.23±1.66	65.83±1.69	+21.40	71.35±1.75	41.07±1.31	+30.28	83.67±0.39	73.35±0.56	+10.33	73.53±0.35	49.07±0.12	+24.46
GCN	88.94±1.98	80.37±1.90	+8.58	80.69±0.96	68.30±1.51	+12.39	88.99±0.57	85.79±0.58	+3.20	78.63±0.21	62.75±0.13	+15.88
GAT	**89.00**±1.90	81.20±1.15	+7.80	80.65±1.26	67.72±1.77	+12.93	88.60±0.44	84.27±0.21	+4.33	78.61±0.16	61.88±0.13	+16.72
SAGE	87.13±2.20	80.16±0.96	+6.97	81.05±0.60	67.39±1.86	+13.66	**90.27**±0.58	88.23±0.37	+2.04	79.10±0.29	61.86±0.23	+17.24
BGRL	88.87±1.79	**82.21**±0.91	+6.66	80.34±0.89	**70.16**±0.99	**+10.18**	89.42±0.55	86.80±0.55	+2.62	79.36±0.21	**64.58**±0.85	**+14.78**
GraphMAE	88.92±2.06	81.67±0.95	+7.25	80.74±1.47	69.51±1.01	+11.23	88.80±0.44	85.07±0.31	+3.73	**79.40**±0.20	63.58±0.44	+15.82
UniGraph2	85.80±2.57	78.07±1.63	+7.73	78.19±1.59	66.83±1.47	+11.37	87.72±0.63	84.51±0.47	+3.21	77.68±0.26	60.53±0.25	+17.16
GIT	87.56±1.39	81.31±1.31	+6.24	**81.47**±0.65	68.40±1.20	+13.07	89.62±0.31	86.84±0.17	+2.78	78.79±0.24	62.47±0.13	+16.32
GFT	86.92±1.76	81.90±1.33	**+5.03**	80.60±1.69	68.32±1.25	+12.28	90.26±0.74	**88.82**±0.48	**+1.44**	75.91±0.38	58.01±0.44	+17.91
OFA	87.08±2.81	77.93±1.32	+9.15	78.18±2.13	60.88±2.06	+17.30	85.29±0.93	81.06±1.09	+4.24	74.06±0.51	59.70±0.23	+14.36
LLaGA	84.35±3.80	76.57±1.40	+7.78	78.37±1.85	60.47±2.36	+17.90	86.86±1.74	81.66±0.73	+5.20		–	

Method	arxiv23			dblp			wikics			elecomp		
	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap
DeepWalk	32.83±0.84	19.08±0.45	+13.74	76.08±0.64	44.45±1.11	+31.63	82.41±1.25	61.64±0.55	+20.77	78.86±0.42	62.95±0.15	+15.91
Node2Vec	32.61±0.94	19.01±0.15	+13.60	76.23±0.54	43.93±1.15	+32.30	83.16±0.82	61.63±1.62	+21.53	79.01±0.47	62.72±0.14	+16.29
GCN	71.17±0.56	66.88±0.21	+4.29	79.66±1.13	75.24±1.81	**+4.42**	85.74±0.86	68.41±3.01	+17.33	86.91±0.40	68.37±0.42	+18.54
GAT	71.05±0.36	66.50±0.54	+4.55	79.64±0.82	71.77±3.33	+7.88	85.99±0.63	72.24±1.77	+13.75	86.66±0.29	67.19±0.24	+19.47
SAGE	72.86±0.48	67.75±0.22	+5.11	81.76±0.85	**76.54**±1.70	+5.22	86.49±0.47	76.15±1.00	+10.34	87.35±0.35	68.82±0.35	+18.53
BGRL	72.57±0.54	**68.76**±0.25	**+3.81**	79.50±1.12	73.25±4.08	+6.25	**87.10**±0.82	65.79±3.48	+21.31	87.34±0.43	68.84±0.56	+18.51
GraphMAE	**72.96**±0.36	68.13±0.18	+4.83	79.99±1.05	75.54±0.68	+4.46	85.64±0.70	**76.51**±0.84	**+9.13**	**88.34**±0.33	**70.12**±0.18	+18.22
UniGraph2	67.12±0.69	60.45±0.49	+6.67	79.22±1.07	74.52±0.89	+4.70	83.35±0.61	71.80±0.89	+11.55	85.68±0.32	64.38±0.62	+21.29
GIT	71.73±0.50	66.31±0.14	+5.42	81.78±1.11	76.24±1.19	+5.54	85.87±0.49	73.84±0.83	+12.03	87.46±0.40	69.15±0.38	+18.30
GFT	70.56±0.65	64.47±0.44	+6.09	**81.81**±0.81	74.43±0.62	+7.38	86.75±0.84	72.81±2.02	+13.94	85.94±0.18	67.92±0.55	+18.02
OFA	65.44±0.70	55.97±0.43	+9.47	79.03±0.77	70.55±2.16	+8.48	83.42±1.23	68.65±0.61	+14.77	87.02±0.66	74.51±0.40	+12.51
LLaGA	69.90±2.89	65.13±1.18	+4.77	76.92±0.61	65.50±1.88	+11.42	82.54±1.52	67.68±1.64	+14.86		–	

Method	elephoto			sportsfit			amazonratings			bookchild		
	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap	ID	OOD	Gap
DeepWalk	74.93±0.71	60.20±0.23	+14.72	88.84±0.17	63.61±0.23	+25.23	38.39±0.50	37.29±0.28	+1.10	47.97±0.40	32.20±0.44	+15.77
Node2Vec	75.14±0.37	60.71±0.30	+14.43	88.97±0.18	63.69±0.13	+25.28	38.31±0.63	36.99±0.46	+1.32	47.96±0.28	32.26±0.47	+15.70
GCN	84.03±0.34	70.28±0.76	+13.76	92.33±0.20	77.77±0.22	+14.56	**43.11**±0.85	43.59±0.59	-0.47	57.33±0.39	43.41±0.46	+13.92
GAT	83.87±0.51	68.29±0.94	+15.59	91.54±0.16	75.17±0.12	+16.37	40.21±0.90	41.08±1.29	-0.87	57.75±0.34	43.59±0.22	+14.16
SAGE	84.76±0.46	71.50±0.76	+13.26	92.91±0.14	77.47±0.20	+15.44	41.27±1.13	44.15±2.97	-2.89	**60.85**±0.33	45.47±0.53	+15.38
BGRL	84.84±0.23	71.49±0.53	+13.35	92.95±0.15	78.25±0.24	+14.70	42.59±0.96	43.34±0.68	-0.75	58.66±0.41	45.49±0.76	+13.17
GraphMAE	**85.72**±0.24	**71.88**±0.28	+13.84	**93.48**±0.15	**80.51**±0.21	**+12.97**	42.74±0.72	43.70±0.87	-0.96	59.52±0.32	45.21±0.32	+14.31
UniGraph2	82.96±0.47	67.86±0.78	+15.10	91.63±0.21	74.27±0.31	+17.36	41.54±0.71	41.98±0.87	-0.43	56.25±0.45	43.28±0.42	**+12.97**
GIT	84.58±0.57	70.89±0.80	+13.68	92.91±0.15	77.12±0.18	+15.80	40.14±0.91	**45.24**±1.32	-5.10	60.63±0.18	**45.86**±0.37	+14.76
GFT	83.32±0.45	68.26±0.65	+15.05	92.35±0.21	75.68±0.40	+16.66	40.86±1.00	41.82±1.58	-0.96	55.23±0.47	40.19±0.54	+15.04
OFA	84.18±0.48	72.08±0.58	+12.11	87.14±0.34	68.65±0.64	+18.49	37.10±0.64	36.16±1.01	+0.94	41.95±0.87	32.46±1.31	+9.49
LLaGA		–			–		39.60±1.95	39.93±2.27	-0.33		–	

Method	bookhis		
	ID	OOD	Gap
DeepWalk	79.18±0.59	69.43±0.25	+9.75
Node2Vec	78.96±0.52	69.10±0.50	+9.85
GCN	83.61±0.51	81.51±0.57	+2.09
GAT	83.67±0.34	80.60±0.98	+3.07
SAGE	**84.79**±0.45	**82.64**±0.24	+2.15
BGRL	84.17±0.66	82.26±0.38	**+1.91**
GraphMAE	84.40±0.59	82.31±0.24	+2.08
UniGraph2	83.04±0.63	80.10±0.43	+2.94
GIT	84.29±0.43	82.25±0.31	+2.04
GFT	83.33±0.49	80.48±0.46	+2.85
OFA	79.91±0.83	75.89±0.76	+4.02
LLaGA	82.40±1.95	80.26±0.40	+2.14

G.3 CLASS IMBALANCE — FULL RESULTS

Table 13: Per-class recall on 8 NC datasets (%): *major / minor* at imbalance ratios $\rho \in \{5, 10, 20\}$, mean±std over 5 seeds. Major and minor groups are ranked independently; **bold /** underline mark the best and second-best methods in each column. Cells marked '–' indicate unavailable results.

Method	cora			citeseer		
	ρ=5	ρ=10	ρ=20	ρ=5	ρ=10	ρ=20
DeepWalk	84.0$_{\pm 0.6}$/24.4$_{\pm 4.2}$	84.9$_{\pm 1.0}$/7.1$_{\pm 2.6}$	85.0$_{\pm 0.9}$/1.9$_{\pm 1.7}$	53.2$_{\pm 1.1}$/9.9$_{\pm 3.5}$	53.7$_{\pm 1.5}$/1.6$_{\pm 0.7}$	54.1$_{\pm 1.3}$/0.2$_{\pm 0.1}$
Node2vec	84.7$_{\pm 0.6}$/23.4$_{\pm 4.1}$	85.7$_{\pm 0.6}$/7.0$_{\pm 2.1}$	85.8$_{\pm 0.8}$/1.9$_{\pm 1.5}$	53.5$_{\pm 1.9}$/9.9$_{\pm 3.4}$	54.4$_{\pm 2.1}$/1.6$_{\pm 0.7}$	54.3$_{\pm 2.0}$/0.3$_{\pm 0.2}$
GCN	85.8$_{\pm 2.0}$/58.3$_{\pm 4.4}$	87.9$_{\pm 1.3}$/41.3$_{\pm 8.2}$	89.9$_{\pm 1.0}$/24.7$_{\pm 10.2}$	73.7$_{\pm 1.8}$/41.1$_{\pm 4.9}$	75.4$_{\pm 2.3}$/28.3$_{\pm 8.6}$	76.9$_{\pm 1.2}$/17.5$_{\pm 6.6}$
GAT	83.8$_{\pm 1.6}$/62.3$_{\pm 5.0}$	85.1$_{\pm 2.9}$/47.4$_{\pm 9.4}$	88.2$_{\pm 2.0}$/27.6$_{\pm 12.6}$	72.5$_{\pm 2.1}$/45.0$_{\pm 5.8}$	73.5$_{\pm 2.0}$/**34.2**$_{\pm 8.2}$	76.3$_{\pm 1.8}$/18.2$_{\pm 9.1}$
GraphSAGE	86.8$_{\pm 1.3}$/49.4$_{\pm 4.5}$	86.7$_{\pm 1.2}$/38.3$_{\pm 5.3}$	87.9$_{\pm 0.8}$/24.2$_{\pm 6.6}$	74.6$_{\pm 1.1}$/38.9$_{\pm 3.0}$	73.9$_{\pm 1.5}$/33.5$_{\pm 4.4}$	75.2$_{\pm 1.3}$/23.8$_{\pm 5.3}$
BGRL	88.2$_{\pm 0.9}$/50.5$_{\pm 4.2}$	89.8$_{\pm 0.8}$/35.4$_{\pm 4.9}$	91.0$_{\pm 0.8}$/20.6$_{\pm 9.0}$	73.9$_{\pm 1.7}$/36.4$_{\pm 3.7}$	75.8$_{\pm 1.0}$/23.4$_{\pm 6.4}$	77.6$_{\pm 1.0}$/11.4$_{\pm 4.3}$
GraphMAE	**88.4**$_{\pm 1.2}$/56.8$_{\pm 6.0}$	**90.3**$_{\pm 0.7}$/40.7$_{\pm 7.8}$	**91.0**$_{\pm 0.8}$/19.7$_{\pm 11.9}$	**77.2**$_{\pm 1.4}$/36.4$_{\pm 4.2}$	**79.0**$_{\pm 1.3}$/21.8$_{\pm 7.6}$	**80.6**$_{\pm 0.9}$/9.2$_{\pm 4.9}$
UniGraph2	77.8$_{\pm 4.4}$/55.5$_{\pm 7.5}$	76.9$_{\pm 7.0}$/42.3$_{\pm 5.5}$	81.2$_{\pm 3.3}$/23.1$_{\pm 9.2}$	63.9$_{\pm 5.8}$/**45.6**$_{\pm 10.8}$	69.0$_{\pm 4.1}$/32.1$_{\pm 5.7}$	68.3$_{\pm 4.3}$/**24.9**$_{\pm 5.0}$
GIT	87.8$_{\pm 1.2}$/44.3$_{\pm 5.5}$	88.8$_{\pm 1.5}$/29.4$_{\pm 5.2}$	89.1$_{\pm 0.5}$/17.6$_{\pm 5.1}$	75.3$_{\pm 1.0}$/36.3$_{\pm 2.9}$	75.3$_{\pm 1.4}$/29.4$_{\pm 5.9}$	77.1$_{\pm 1.0}$/19.1$_{\pm 5.2}$
GFT	84.6$_{\pm 1.6}$/53.6$_{\pm 5.5}$	87.6$_{\pm 1.6}$/36.1$_{\pm 6.6}$	88.1$_{\pm 1.8}$/20.7$_{\pm 6.2}$	73.6$_{\pm 2.1}$/39.7$_{\pm 5.5}$	75.6$_{\pm 1.8}$/28.9$_{\pm 5.3}$	77.1$_{\pm 1.5}$/18.4$_{\pm 3.9}$
OFA	83.0$_{\pm 0.2}$/47.3$_{\pm 8.2}$	84.2$_{\pm 0.8}$/34.7$_{\pm 8.0}$	85.5$_{\pm 0.5}$/18.0$_{\pm 6.6}$	68.1$_{\pm 8.2}$/36.7$_{\pm 8.0}$	69.3$_{\pm 3.1}$/25.5$_{\pm 4.1}$	73.7$_{\pm 2.5}$/10.5$_{\pm 7.5}$
LLaGA	83.6$_{\pm 2.0}$/**63.9**$_{\pm 7.9}$	85.4$_{\pm 1.1}$/**54.1**$_{\pm 10.0}$	87.8$_{\pm 1.3}$/**37.1**$_{\pm 8.0}$	74.0$_{\pm 2.8}$/29.2$_{\pm 8.2}$	77.4$_{\pm 2.1}$/10.8$_{\pm 9.5}$	75.3$_{\pm 3.2}$/8.3$_{\pm 5.7}$

Method	pubmed			dblp		
	ρ=5	ρ=10	ρ=20	ρ=5	ρ=10	ρ=20
DeepWalk	**99.5**$_{\pm 0.3}$/7.4$_{\pm 1.4}$	**99.9**$_{\pm 0.1}$/1.5$_{\pm 1.2}$	**100.0**$_{\pm 0.0}$/0.1$_{\pm 0.2}$	65.6$_{\pm 1.0}$/3.8$_{\pm 1.5}$	66.0$_{\pm 1.1}$/0.6$_{\pm 0.2}$	66.0$_{\pm 1.0}$/0.0$_{\pm 0.1}$
Node2vec	**99.5**$_{\pm 0.4}$/7.3$_{\pm 2.1}$	**99.9**$_{\pm 0.0}$/1.5$_{\pm 1.5}$	**100.0**$_{\pm 0.0}$/0.1$_{\pm 0.2}$	65.3$_{\pm 1.3}$/4.1$_{\pm 2.3}$	65.7$_{\pm 1.4}$/0.7$_{\pm 0.3}$	65.7$_{\pm 1.2}$/0.1$_{\pm 0.1}$
GCN	91.4$_{\pm 4.4}$/50.2$_{\pm 7.5}$	96.8$_{\pm 2.2}$/32.1$_{\pm 7.5}$	99.2$_{\pm 0.5}$/12.4$_{\pm 9.7}$	83.8$_{\pm 1.5}$/20.0$_{\pm 6.0}$	85.5$_{\pm 2.3}$/8.3$_{\pm 7.8}$	86.4$_{\pm 2.2}$/4.8$_{\pm 6.8}$
GAT	86.0$_{\pm 9.1}$/56.5$_{\pm 6.8}$	93.2$_{\pm 7.0}$/39.6$_{\pm 13.4}$	97.5$_{\pm 3.4}$/19.6$_{\pm 16.6}$	84.1$_{\pm 1.7}$/24.9$_{\pm 8.5}$	84.7$_{\pm 3.4}$/12.5$_{\pm 11.5}$	86.5$_{\pm 1.9}$/5.9$_{\pm 9.0}$
GraphSAGE	92.0$_{\pm 4.0}$/44.7$_{\pm 7.3}$	96.7$_{\pm 1.9}$/31.1$_{\pm 9.5}$	99.4$_{\pm 0.6}$/10.1$_{\pm 5.7}$	83.4$_{\pm 1.7}$/22.4$_{\pm 5.8}$	84.1$_{\pm 1.8}$/14.2$_{\pm 5.5}$	84.1$_{\pm 1.4}$/10.9$_{\pm 4.5}$
BGRL	89.0$_{\pm 4.3}$/50.6$_{\pm 5.4}$	93.2$_{\pm 3.1}$/42.0$_{\pm 8.5}$	94.2$_{\pm 3.9}$/33.0$_{\pm 9.4}$	83.3$_{\pm 1.9}$/27.4$_{\pm 6.5}$	85.2$_{\pm 1.7}$/17.1$_{\pm 9.2}$	83.9$_{\pm 6.1}$/11.6$_{\pm 8.7}$
GraphMAE	93.4$_{\pm 3.4}$/39.3$_{\pm 9.2}$	98.3$_{\pm 1.4}$/21.1$_{\pm 7.0}$	98.9$_{\pm 2.7}$/8.3$_{\pm 6.8}$	**84.7**$_{\pm 1.4}$/19.6$_{\pm 7.2}$	**86.5**$_{\pm 1.4}$/7.7$_{\pm 8.3}$	**87.2**$_{\pm 1.2}$/3.1$_{\pm 5.9}$
UniGraph2	79.3$_{\pm 6.3}$/59.3$_{\pm 3.6}$	84.2$_{\pm 4.3}$/52.9$_{\pm 5.1}$	80.0$_{\pm 7.6}$/**49.5**$_{\pm 8.2}$	67.7$_{\pm 11.3}$/**40.0**$_{\pm 13.6}$	71.3$_{\pm 5.4}$/**30.4**$_{\pm 13.3}$	76.1$_{\pm 6.7}$/**14.9**$_{\pm 11.5}$
GIT	92.8$_{\pm 3.4}$/43.8$_{\pm 7.3}$	96.6$_{\pm 2.2}$/30.7$_{\pm 10.4}$	99.1$_{\pm 0.8}$/12.3$_{\pm 7.4}$	84.0$_{\pm 1.8}$/17.6$_{\pm 5.9}$	85.3$_{\pm 1.7}$/8.7$_{\pm 4.6}$	86.6$_{\pm 0.7}$/3.6$_{\pm 2.4}$
GFT	90.0$_{\pm 5.2}$/53.4$_{\pm 8.6}$	96.7$_{\pm 1.8}$/36.2$_{\pm 8.3}$	99.1$_{\pm 0.4}$/16.5$_{\pm 9.9}$	81.4$_{\pm 2.0}$/31.1$_{\pm 8.0}$	82.9$_{\pm 2.2}$/15.2$_{\pm 6.4}$	84.6$_{\pm 1.1}$/8.3$_{\pm 3.5}$
OFA	73.3$_{\pm 1.6}$/**66.2**$_{\pm 7.7}$	72.9$_{\pm 2.2}$/**57.6**$_{\pm 8.5}$	75.4$_{\pm 2.2}$/29.6$_{\pm 21.9}$	69.7$_{\pm 7.0}$/31.2$_{\pm 6.9}$	79.6$_{\pm 5.3}$/12.5$_{\pm 8.5}$	82.3$_{\pm 3.5}$/3.6$_{\pm 7.9}$
LLaGA	86.7$_{\pm 12.9}$/50.5$_{\pm 11.5}$	94.3$_{\pm 4.7}$/34.3$_{\pm 13.5}$	98.5$_{\pm 1.6}$/21.4$_{\pm 9.9}$	80.9$_{\pm 1.0}$/22.3$_{\pm 9.0}$	85.8$_{\pm 2.5}$/4.1$_{\pm 7.1}$	86.7$_{\pm 1.6}$/0.0$_{\pm 0.0}$

Method	elecomp			elephoto		
	ρ=5	ρ=10	ρ=20	ρ=5	ρ=10	ρ=20
DeepWalk	73.1$_{\pm 0.5}$/54.0$_{\pm 3.5}$	73.9$_{\pm 0.5}$/49.0$_{\pm 3.1}$	74.3$_{\pm 0.5}$/44.3$_{\pm 2.8}$	65.4$_{\pm 0.2}$/55.1$_{\pm 1.1}$	65.5$_{\pm 0.3}$/52.9$_{\pm 1.4}$	65.9$_{\pm 0.4}$/46.0$_{\pm 0.7}$
Node2vec	73.1$_{\pm 0.4}$/52.2$_{\pm 2.1}$	73.9$_{\pm 0.4}$/48.6$_{\pm 1.6}$	74.2$_{\pm 0.4}$/43.3$_{\pm 1.9}$	65.2$_{\pm 0.4}$/55.8$_{\pm 1.5}$	65.4$_{\pm 0.3}$/54.1$_{\pm 1.5}$	65.7$_{\pm 0.4}$/47.4$_{\pm 1.5}$
GCN	77.6$_{\pm 0.3}$/55.0$_{\pm 1.3}$	78.0$_{\pm 0.2}$/51.1$_{\pm 0.9}$	78.2$_{\pm 0.3}$/46.8$_{\pm 0.8}$	77.0$_{\pm 0.5}$/57.7$_{\pm 1.3}$	76.8$_{\pm 0.4}$/56.7$_{\pm 0.6}$	77.4$_{\pm 0.5}$/54.0$_{\pm 1.2}$
GAT	76.3$_{\pm 0.1}$/54.6$_{\pm 0.3}$	76.9$_{\pm 0.2}$/51.8$_{\pm 1.0}$	77.2$_{\pm 0.1}$/48.3$_{\pm 0.5}$	76.1$_{\pm 0.5}$/55.3$_{\pm 0.9}$	76.2$_{\pm 0.4}$/54.0$_{\pm 1.0}$	76.8$_{\pm 0.5}$/50.3$_{\pm 1.4}$
GraphSAGE	76.7$_{\pm 0.4}$/54.3$_{\pm 1.1}$	77.4$_{\pm 0.2}$/51.5$_{\pm 0.7}$	77.6$_{\pm 0.5}$/47.7$_{\pm 0.9}$	76.3$_{\pm 0.7}$/56.9$_{\pm 1.2}$	76.0$_{\pm 0.5}$/55.2$_{\pm 1.0}$	76.7$_{\pm 0.4}$/50.9$_{\pm 1.1}$
BGRL	75.4$_{\pm 0.4}$/56.0$_{\pm 1.6}$	75.9$_{\pm 0.6}$/54.8$_{\pm 1.5}$	76.3$_{\pm 0.6}$/50.3$_{\pm 2.3}$	75.8$_{\pm 0.5}$/56.1$_{\pm 1.5}$	76.0$_{\pm 0.3}$/53.9$_{\pm 1.3}$	76.7$_{\pm 0.6}$/50.6$_{\pm 1.4}$
GraphMAE	78.8$_{\pm 0.2}$/58.9$_{\pm 0.8}$	79.1$_{\pm 0.4}$/56.6$_{\pm 1.1}$	79.5$_{\pm 0.3}$/53.4$_{\pm 0.9}$	77.2$_{\pm 0.2}$/57.8$_{\pm 0.4}$	77.5$_{\pm 0.2}$/56.6$_{\pm 0.5}$	78.0$_{\pm 0.2}$/53.4$_{\pm 0.5}$
UniGraph2	75.1$_{\pm 0.7}$/**59.2**$_{\pm 2.5}$	75.0$_{\pm 1.1}$/**58.7**$_{\pm 3.2}$	75.6$_{\pm 1.4}$/**57.3**$_{\pm 2.5}$	75.1$_{\pm 1.3}$/**60.1**$_{\pm 2.1}$	75.1$_{\pm 1.3}$/**59.0**$_{\pm 1.5}$	75.9$_{\pm 1.3}$/**56.0**$_{\pm 2.4}$
GIT	77.0$_{\pm 0.3}$/54.2$_{\pm 0.9}$	77.7$_{\pm 0.4}$/51.7$_{\pm 1.2}$	77.9$_{\pm 0.4}$/48.5$_{\pm 1.2}$	76.4$_{\pm 0.7}$/57.7$_{\pm 1.0}$	76.5$_{\pm 0.9}$/56.4$_{\pm 0.9}$	77.0$_{\pm 0.8}$/50.9$_{\pm 0.9}$
GFT	77.4$_{\pm 0.3}$/56.9$_{\pm 1.5}$	77.8$_{\pm 0.2}$/54.6$_{\pm 1.4}$	78.0$_{\pm 0.3}$/51.5$_{\pm 1.5}$	**77.8**$_{\pm 0.6}$/59.3$_{\pm 0.9}$	**77.9**$_{\pm 0.4}$/58.5$_{\pm 1.6}$	**78.1**$_{\pm 0.5}$/53.4$_{\pm 1.3}$
OFA	77.0$_{\pm 0.1}$/36.3$_{\pm 0.5}$	77.5$_{\pm 0.1}$/29.2$_{\pm 0.7}$	77.5$_{\pm 0.1}$/20.5$_{\pm 1.0}$	75.1$_{\pm 0.1}$/28.3$_{\pm 0.2}$	75.2$_{\pm 0.4}$/18.4$_{\pm 1.3}$	75.2$_{\pm 0.3}$/8.3$_{\pm 0.8}$
LLaGA	**81.1**$_{\pm 2.8}$/46.9$_{\pm 4.5}$	**81.1**$_{\pm 2.8}$/41.2$_{\pm 8.4}$	**81.8**$_{\pm 2.2}$/41.8$_{\pm 8.3}$	75.4$_{\pm 6.6}$/55.2$_{\pm 8.5}$	75.3$_{\pm 6.5}$/56.3$_{\pm 12.2}$	74.5$_{\pm 6.9}$/46.6$_{\pm 7.7}$

Method	sportsfit			wikics		
	ρ=5	ρ=10	ρ=20	ρ=5	ρ=10	ρ=20
DeepWalk	82.1$_{\pm 0.2}$/45.0$_{\pm 0.2}$	82.2$_{\pm 0.2}$/44.9$_{\pm 0.2}$	83.0$_{\pm 0.2}$/42.0$_{\pm 0.2}$	76.0$_{\pm 0.3}$/58.0$_{\pm 1.9}$	77.3$_{\pm 0.5}$/44.0$_{\pm 1.2}$	78.0$_{\pm 0.5}$/20.9$_{\pm 2.0}$
Node2vec	82.1$_{\pm 0.1}$/45.5$_{\pm 0.3}$	82.2$_{\pm 0.1}$/45.4$_{\pm 0.3}$	83.0$_{\pm 0.1}$/42.6$_{\pm 0.3}$	75.3$_{\pm 0.5}$/60.2$_{\pm 1.8}$	76.6$_{\pm 0.6}$/44.3$_{\pm 3.8}$	77.3$_{\pm 0.6}$/20.6$_{\pm 2.4}$
GCN	89.7$_{\pm 0.2}$/72.8$_{\pm 0.7}$	89.6$_{\pm 0.2}$/73.3$_{\pm 1.3}$	90.1$_{\pm 0.2}$/70.1$_{\pm 1.3}$	79.9$_{\pm 0.3}$/69.7$_{\pm 1.4}$	81.3$_{\pm 0.8}$/62.2$_{\pm 2.9}$	81.1$_{\pm 0.7}$/49.3$_{\pm 2.9}$
GAT	87.4$_{\pm 0.2}$/67.8$_{\pm 0.4}$	87.4$_{\pm 0.2}$/67.9$_{\pm 0.7}$	87.9$_{\pm 0.2}$/65.0$_{\pm 0.5}$	79.6$_{\pm 0.4}$/72.3$_{\pm 1.3}$	80.3$_{\pm 0.6}$/65.2$_{\pm 2.4}$	81.0$_{\pm 0.7}$/**53.0**$_{\pm 3.3}$
GraphSAGE	88.9$_{\pm 0.1}$/72.8$_{\pm 0.4}$	89.0$_{\pm 0.1}$/72.4$_{\pm 0.5}$	89.4$_{\pm 0.1}$/69.5$_{\pm 0.5}$	78.9$_{\pm 0.3}$/70.2$_{\pm 0.8}$	80.2$_{\pm 0.4}$/63.3$_{\pm 2.0}$	81.3$_{\pm 0.3}$/48.0$_{\pm 1.6}$
BGRL	88.5$_{\pm 0.5}$/**74.0**$_{\pm 0.8}$	88.5$_{\pm 0.5}$/73.9$_{\pm 0.6}$	89.1$_{\pm 0.3}$/70.9$_{\pm 0.6}$	**81.8**$_{\pm 0.3}$/68.8$_{\pm 1.3}$	**82.9**$_{\pm 0.3}$/59.1$_{\pm 1.6}$	**83.0**$_{\pm 0.6}$/43.5$_{\pm 3.4}$
GraphMAE	**90.7**$_{\pm 0.1}$/73.5$_{\pm 0.4}$	**90.7**$_{\pm 0.1}$/73.4$_{\pm 0.4}$	**91.1**$_{\pm 0.1}$/70.8$_{\pm 0.5}$	80.8$_{\pm 0.5}$/**73.6**$_{\pm 1.0}$	81.7$_{\pm 0.4}$/**67.9**$_{\pm 1.5}$	82.7$_{\pm 0.2}$/52.7$_{\pm 2.1}$
UniGraph2	84.5$_{\pm 1.6}$/73.3$_{\pm 2.1}$	84.8$_{\pm 1.2}$/71.3$_{\pm 3.0}$	85.7$_{\pm 0.6}$/69.5$_{\pm 1.9}$	74.0$_{\pm 1.3}$/67.6$_{\pm 2.2}$	76.3$_{\pm 1.3}$/59.9$_{\pm 3.8}$	76.1$_{\pm 1.7}$/47.7$_{\pm 2.7}$
GIT	89.3$_{\pm 0.2}$/73.4$_{\pm 0.4}$	89.3$_{\pm 0.2}$/73.2$_{\pm 0.3}$	89.6$_{\pm 0.2}$/70.2$_{\pm 0.7}$	78.8$_{\pm 0.6}$/70.5$_{\pm 1.1}$	79.9$_{\pm 0.4}$/62.8$_{\pm 1.5}$	81.3$_{\pm 0.2}$/46.5$_{\pm 2.1}$
GFT	89.6$_{\pm 0.1}$/73.9$_{\pm 0.7}$	89.5$_{\pm 0.2}$/**74.0**$_{\pm 0.8}$	89.9$_{\pm 0.3}$/**71.4**$_{\pm 1.3}$	78.4$_{\pm 0.5}$/71.5$_{\pm 0.7}$	79.4$_{\pm 0.7}$/64.2$_{\pm 1.6}$	80.8$_{\pm 0.2}$/49.9$_{\pm 3.1}$
OFA	–	–	–	78.5$_{\pm 0.6}$/49.3$_{\pm 3.1}$	79.1$_{\pm 0.3}$/36.4$_{\pm 4.4}$	79.7$_{\pm 0.3}$/20.2$_{\pm 6.5}$
LLaGA	84.8$_{\pm 2.6}$/41.9$_{\pm 12.0}$	85.7$_{\pm 2.4}$/43.7$_{\pm 13.4}$	84.3$_{\pm 1.8}$/35.0$_{\pm 10.6}$	76.0$_{\pm 1.8}$/62.7$_{\pm 10.0}$	77.6$_{\pm 2.2}$/51.5$_{\pm 7.4}$	78.8$_{\pm 1.7}$/40.8$_{\pm 9.0}$

G.4 FAIRNESS — FULL PER-DATASET RESULTS

The main-text fairness section reports structural fairness. Full per-dataset structural-fairness results are reported below for eight NC datasets, including the four representative datasets shown in Figure 5. Demographic fairness is evaluated on tolokers (banned-user prediction; sensitive attribute: education) and reported in Table 14.

Table 14: **Demographic fairness.** Test AUC and the three fairness gaps on Tolokers (banned-user prediction; sensitive attribute: education). Higher AUC is better; lower Δ is more fair. All values in % (Δ values $\times 100$). Mean$\pm$std over 5 seeds. **Bold** / underline = best / second AUC. $\Delta_{\text{SP}}=|P(\hat{Y}=1|S=0)-P(\hat{Y}=1|S=1)|$, $\Delta_{\text{EO}}=|\text{TPR}_{S=0}-\text{TPR}_{S=1}|$, $\Delta_{\text{Util}}=|\text{AUC}_{S=0}-\text{AUC}_{S=1}|$.

Method	AUC↑	Δ_{SP}↓	Δ_{EO}↓	Δ_{Util}↓
GCN	$74.78_{\pm 0.31}$	$9.02_{\pm 0.73}$	$11.80_{\pm 0.81}$	$10.16_{\pm 0.41}$
GAT	$70.01_{\pm 2.27}$	$1.24_{\pm 1.53}$	$2.31_{\pm 1.76}$	$7.78_{\pm 1.69}$
SAGE	$75.59_{\pm 0.45}$	$6.59_{\pm 2.52}$	$9.21_{\pm 2.59}$	$13.65_{\pm 0.34}$
BGRL	$\mathbf{80.41}_{\pm 0.36}$	$9.65_{\pm 1.74}$	$13.33_{\pm 1.88}$	$8.30_{\pm 0.80}$
GraphMAE	$68.90_{\pm 3.46}$	$1.12_{\pm 1.48}$	$1.89_{\pm 2.44}$	$11.73_{\pm 4.45}$
GFT	$72.42_{\pm 0.61}$	$1.24_{\pm 0.74}$	$2.03_{\pm 1.15}$	$11.43_{\pm 0.77}$
GIT	$\underline{76.66}_{\pm 0.44}$	$6.86_{\pm 1.88}$	$9.77_{\pm 2.48}$	$14.06_{\pm 0.31}$
UniGraph2	$69.34_{\pm 5.20}$	$2.14_{\pm 3.40}$	$3.53_{\pm 5.52}$	$9.12_{\pm 4.01}$

Table 15: Full **structural-fairness** per-dataset results, one block per dataset. Columns: overall accuracy, accuracy on head (high-degree) nodes, accuracy on tail (low-degree) nodes, and head−tail gap. Within each dataset block, **bold** = method with smallest |gap|; underline = second. Cells report mean$_{\pm \text{std}}$ over 5 seeds.

Cora

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$71.47_{\pm 1.60}$	$78.83_{\pm 0.72}$	$63.12_{\pm 3.05}$	$+15.71_{\pm 3.28}$
Node2vec	$71.45_{\pm 1.39}$	$79.14_{\pm 1.26}$	$62.84_{\pm 2.21}$	$+16.30_{\pm 2.29}$
GCN	$79.34_{\pm 0.66}$	$81.73_{\pm 1.58}$	$76.02_{\pm 0.62}$	$+5.71_{\pm 2.08}$
GAT	$79.20_{\pm 0.96}$	$81.75_{\pm 1.24}$	$76.06_{\pm 1.02}$	$+5.68_{\pm 0.97}$
SAGE	$78.60_{\pm 1.22}$	$80.63_{\pm 2.45}$	$75.42_{\pm 0.88}$	$+5.21_{\pm 2.16}$
BGRL	$76.68_{\pm 1.04}$	$80.14_{\pm 2.02}$	$71.81_{\pm 1.36}$	$+8.34_{\pm 3.05}$
GraphMAE	$79.89_{\pm 1.49}$	$82.33_{\pm 2.32}$	$76.40_{\pm 1.51}$	$+5.93_{\pm 2.52}$
UniGraph2	$75.93_{\pm 1.52}$	$76.14_{\pm 2.33}$	$72.92_{\pm 1.99}$	**+3.22**$_{\pm 3.25}$
GIT	$78.59_{\pm 1.13}$	$80.36_{\pm 1.56}$	$75.39_{\pm 0.91}$	$+4.97_{\pm 1.39}$
GFT	$79.16_{\pm 0.53}$	$82.00_{\pm 1.37}$	$74.83_{\pm 0.59}$	$+7.17_{\pm 1.80}$
OFA	$75.17_{\pm 1.64}$	$77.44_{\pm 2.87}$	$73.70_{\pm 1.10}$	$\underline{+3.74}_{\pm 2.33}$
LLaGA	$76.31_{\pm 1.38}$	$79.19_{\pm 2.58}$	$72.35_{\pm 1.37}$	$+6.84_{\pm 3.02}$

CiteSeer

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$47.04_{\pm 0.66}$	$68.79_{\pm 0.80}$	$33.36_{\pm 0.71}$	$+35.43_{\pm 0.23}$
Node2vec	$46.56_{\pm 0.74}$	$68.07_{\pm 0.66}$	$33.30_{\pm 1.36}$	$+34.76_{\pm 0.96}$
GCN	$68.51_{\pm 0.64}$	$77.35_{\pm 1.26}$	$62.74_{\pm 0.75}$	$+14.60_{\pm 1.48}$
GAT	$69.62_{\pm 0.58}$	$77.54_{\pm 0.53}$	$64.28_{\pm 0.61}$	$+13.26_{\pm 0.51}$
SAGE	$69.16_{\pm 0.40}$	$77.35_{\pm 0.35}$	$63.90_{\pm 0.51}$	$+13.45_{\pm 0.59}$
BGRL	$66.31_{\pm 0.55}$	$76.93_{\pm 0.84}$	$58.95_{\pm 0.88}$	$+17.98_{\pm 0.79}$
GraphMAE	$67.92_{\pm 0.46}$	$78.52_{\pm 0.49}$	$61.23_{\pm 1.04}$	$+17.30_{\pm 0.97}$
UniGraph2	$68.08_{\pm 1.10}$	$75.64_{\pm 3.32}$	$63.16_{\pm 1.30}$	$\underline{+12.49}_{\pm 3.88}$
GIT	$69.31_{\pm 0.54}$	$77.12_{\pm 0.47}$	$64.19_{\pm 0.54}$	$+12.93_{\pm 0.49}$
GFT	$66.77_{\pm 0.45}$	$78.48_{\pm 0.89}$	$59.57_{\pm 1.19}$	$+18.92_{\pm 1.98}$
OFA	$63.83_{\pm 1.01}$	$70.02_{\pm 1.43}$	$60.79_{\pm 0.93}$	**+9.23**$_{\pm 0.99}$
LLaGA	$64.12_{\pm 1.04}$	$74.37_{\pm 2.14}$	$56.42_{\pm 2.05}$	$+17.95_{\pm 3.72}$

PubMed

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$67.73_{\pm 2.18}$	$76.82_{\pm 1.73}$	$61.12_{\pm 2.44}$	$+15.71_{\pm 1.23}$
Node2vec	$67.42_{\pm 2.60}$	$76.47_{\pm 3.03}$	$60.79_{\pm 2.56}$	$+15.68_{\pm 1.14}$
GCN	$78.45_{\pm 1.75}$	$79.49_{\pm 2.08}$	$76.92_{\pm 1.64}$	$\underline{+2.56}_{\pm 0.73}$
GAT	$78.49_{\pm 0.87}$	$80.51_{\pm 0.81}$	$76.43_{\pm 1.05}$	$+4.08_{\pm 0.64}$
SAGE	$76.20_{\pm 0.99}$	$78.46_{\pm 1.21}$	$73.55_{\pm 1.00}$	$+4.91_{\pm 0.90}$
BGRL	$73.10_{\pm 2.83}$	$72.22_{\pm 4.11}$	$72.51_{\pm 2.56}$	**−0.28**$_{\pm 2.07}$
GraphMAE	$77.92_{\pm 1.75}$	$80.39_{\pm 2.23}$	$75.54_{\pm 1.85}$	$+4.85_{\pm 0.92}$
UniGraph2	$75.43_{\pm 1.37}$	$77.66_{\pm 2.24}$	$73.19_{\pm 1.38}$	$+4.47_{\pm 2.43}$
GIT	$76.29_{\pm 0.76}$	$78.80_{\pm 0.97}$	$73.57_{\pm 1.05}$	$+5.23_{\pm 1.31}$
GFT	$78.03_{\pm 0.56}$	$80.55_{\pm 1.30}$	$75.19_{\pm 0.70}$	$+5.36_{\pm 1.46}$
OFA	$73.41_{\pm 2.56}$	$75.99_{\pm 3.22}$	$71.89_{\pm 2.20}$	$+4.10_{\pm 1.41}$
LLaGA	$73.56_{\pm 2.63}$	$75.67_{\pm 3.64}$	$70.94_{\pm 2.38}$	$+4.73_{\pm 1.47}$

WikiCS

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$73.49_{\pm 0.96}$	$81.81_{\pm 1.17}$	$61.72_{\pm 0.94}$	$+20.09_{\pm 1.00}$
Node2vec	$73.31_{\pm 0.54}$	$82.25_{\pm 1.09}$	$61.36_{\pm 0.98}$	$+20.89_{\pm 0.85}$
GCN	$78.81_{\pm 0.73}$	$84.14_{\pm 1.42}$	$73.29_{\pm 0.74}$	$+10.85_{\pm 1.62}$
GAT	$78.93_{\pm 0.48}$	$84.18_{\pm 1.56}$	$73.15_{\pm 0.45}$	$+11.03_{\pm 1.54}$
SAGE	$78.46_{\pm 0.34}$	$84.83_{\pm 1.19}$	$72.13_{\pm 0.51}$	$+12.70_{\pm 1.23}$
BGRL	$79.25_{\pm 0.53}$	$83.97_{\pm 1.47}$	$74.26_{\pm 1.07}$	$+9.70_{\pm 1.26}$
GraphMAE	$79.82_{\pm 0.55}$	$84.59_{\pm 1.18}$	$75.07_{\pm 0.94}$	$+9.51_{\pm 1.28}$
UniGraph2	$74.94_{\pm 1.09}$	$78.70_{\pm 2.19}$	$70.09_{\pm 1.19}$	$\underline{+8.61}_{\pm 2.13}$
GIT	$78.74_{\pm 0.42}$	$84.85_{\pm 1.01}$	$72.36_{\pm 0.59}$	$+12.49_{\pm 1.00}$
GFT	$78.51_{\pm 0.58}$	$84.74_{\pm 0.78}$	$70.16_{\pm 1.44}$	$+14.58_{\pm 1.25}$
OFA	$76.55_{\pm 0.68}$	$80.16_{\pm 0.88}$	$73.23_{\pm 0.67}$	**+6.93**$_{\pm 0.52}$
LLaGA	$72.84_{\pm 1.01}$	$78.03_{\pm 1.64}$	$68.33_{\pm 3.59}$	$+9.70_{\pm 4.45}$

ElecComp

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$71.10_{\pm 0.39}$	$78.70_{\pm 0.83}$	$57.74_{\pm 0.82}$	$+20.96_{\pm 1.46}$
Node2vec	$71.00_{\pm 0.09}$	$78.27_{\pm 0.33}$	$57.55_{\pm 0.65}$	$+20.72_{\pm 0.84}$
GCN	$76.79_{\pm 0.13}$	$86.78_{\pm 0.20}$	$66.47_{\pm 0.37}$	$+20.31_{\pm 0.46}$
GAT	$75.20_{\pm 0.17}$	$85.75_{\pm 0.32}$	$64.30_{\pm 0.36}$	$+21.45_{\pm 0.30}$
SAGE	$76.37_{\pm 0.16}$	$86.35_{\pm 0.33}$	$65.74_{\pm 0.24}$	$+20.61_{\pm 0.32}$
BGRL	$76.32_{\pm 0.12}$	$86.22_{\pm 0.22}$	$65.41_{\pm 0.25}$	$+20.81_{\pm 0.41}$
GraphMAE	$78.06_{\pm 0.16}$	$87.38_{\pm 0.26}$	$67.64_{\pm 0.46}$	$+19.75_{\pm 0.37}$
UniGraph2	$75.47_{\pm 0.26}$	$85.93_{\pm 0.11}$	$64.18_{\pm 0.71}$	$+21.75_{\pm 0.72}$
GIT	$76.54_{\pm 0.17}$	$86.48_{\pm 0.23}$	$66.20_{\pm 0.29}$	$+20.28_{\pm 0.35}$
GFT	$77.36_{\pm 0.15}$	$87.03_{\pm 0.29}$	$66.86_{\pm 0.48}$	$+20.17_{\pm 0.55}$
OFA	$78.08_{\pm 1.11}$	$83.23_{\pm 1.00}$	$73.32_{\pm 1.17}$	**+9.91**$_{\pm 0.50}$
LLaGA	$77.31_{\pm 2.53}$	$81.20_{\pm 3.51}$	$68.78_{\pm 3.96}$	$\underline{+12.42}_{\pm 5.56}$

ElePhoto

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$71.56_{\pm 0.14}$	$74.75_{\pm 0.39}$	$64.80_{\pm 0.46}$	$+9.95_{\pm 0.33}$
Node2vec	$71.71_{\pm 0.22}$	$74.93_{\pm 0.58}$	$65.31_{\pm 0.56}$	$+9.62_{\pm 1.03}$
GCN	$78.42_{\pm 0.13}$	$80.52_{\pm 0.26}$	$76.89_{\pm 0.33}$	$+3.63_{\pm 0.52}$
GAT	$77.75_{\pm 0.14}$	$80.12_{\pm 0.38}$	$75.33_{\pm 0.17}$	$+4.79_{\pm 0.44}$
SAGE	$78.64_{\pm 0.07}$	$80.37_{\pm 0.20}$	$77.74_{\pm 0.19}$	$+2.63_{\pm 0.26}$
BGRL	$78.90_{\pm 0.19}$	$80.96_{\pm 0.39}$	$77.59_{\pm 0.20}$	$+3.37_{\pm 0.43}$
GraphMAE	$79.89_{\pm 0.16}$	$82.53_{\pm 0.19}$	$78.12_{\pm 0.33}$	$+4.42_{\pm 0.30}$
UniGraph2	$77.42_{\pm 0.12}$	$80.07_{\pm 0.62}$	$75.64_{\pm 0.31}$	$+4.43_{\pm 0.68}$
GIT	$78.70_{\pm 0.12}$	$80.13_{\pm 0.25}$	$77.79_{\pm 0.16}$	$\underline{+2.34}_{\pm 0.24}$
GFT	$79.11_{\pm 0.11}$	$81.54_{\pm 0.38}$	$77.92_{\pm 0.22}$	$+3.62_{\pm 0.58}$
OFA	$79.24_{\pm 1.41}$	$80.64_{\pm 1.83}$	$78.13_{\pm 1.22}$	$+2.51_{\pm 0.90}$
LLaGA	$79.03_{\pm 1.18}$	$76.35_{\pm 1.58}$	$77.19_{\pm 2.68}$	**−0.83**$_{\pm 2.94}$

AmazonRatings

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$38.63_{\pm 0.44}$	$38.91_{\pm 0.31}$	$39.36_{\pm 0.65}$	$\underline{-0.45}_{\pm 0.81}$
Node2vec	$38.62_{\pm 0.23}$	$39.16_{\pm 0.70}$	$39.47_{\pm 0.42}$	**−0.31**$_{\pm 0.75}$
GCN	$46.17_{\pm 0.31}$	$43.10_{\pm 0.49}$	$47.43_{\pm 0.44}$	$-4.33_{\pm 0.42}$
GAT	$43.34_{\pm 0.40}$	$39.56_{\pm 0.68}$	$45.67_{\pm 0.28}$	$-6.11_{\pm 0.62}$
SAGE	$49.16_{\pm 0.34}$	$44.58_{\pm 0.46}$	$51.91_{\pm 0.64}$	$-7.33_{\pm 0.78}$
BGRL	$46.43_{\pm 0.07}$	$40.07_{\pm 0.09}$	$48.63_{\pm 0.05}$	$-8.55_{\pm 0.10}$
GraphMAE	$46.63_{\pm 0.00}$	$41.19_{\pm 0.00}$	$49.22_{\pm 0.00}$	$-8.03_{\pm 0.00}$
UniGraph2	$45.16_{\pm 0.00}$	$41.24_{\pm 0.00}$	$47.50_{\pm 0.00}$	$-6.26_{\pm 0.00}$
GIT	$48.78_{\pm 0.04}$	$43.62_{\pm 0.52}$	$51.77_{\pm 0.28}$	$-8.14_{\pm 0.59}$
GFT	$47.43_{\pm 0.22}$	$40.77_{\pm 0.24}$	$50.94_{\pm 0.47}$	$-10.17_{\pm 0.51}$
OFA	$36.56_{\pm 0.94}$	$35.70_{\pm 1.08}$	$37.27_{\pm 1.20}$	$-1.58_{\pm 1.38}$
LLaGA	$37.04_{\pm 2.14}$	$36.70_{\pm 4.30}$	$38.19_{\pm 1.29}$	$-1.49_{\pm 4.70}$

Tolokers

Method	Acc	Acc(h)	Acc(t)	Gap↓
DeepWalk	$80.41_{\pm 0.29}$	$75.78_{\pm 0.57}$	$83.55_{\pm 0.32}$	$-7.77_{\pm 0.62}$
Node2vec	$80.19_{\pm 0.15}$	$75.07_{\pm 1.24}$	$83.64_{\pm 0.39}$	$-8.58_{\pm 1.52}$
GCN	$78.66_{\pm 0.09}$	$74.66_{\pm 0.49}$	$83.42_{\pm 0.10}$	$-8.77_{\pm 0.54}$
GAT	$78.14_{\pm 0.04}$	$72.13_{\pm 0.00}$	$83.40_{\pm 0.18}$	$-11.27_{\pm 0.18}$
SAGE	$78.60_{\pm 0.33}$	$72.32_{\pm 0.45}$	$83.25_{\pm 0.68}$	$-10.92_{\pm 0.62}$
BGRL	$80.42_{\pm 0.33}$	$76.16_{\pm 0.85}$	$82.63_{\pm 1.02}$	$\underline{-6.47}_{\pm 1.26}$
GraphMAE	$78.16_{\pm 0.00}$	$72.13_{\pm 0.00}$	$83.49_{\pm 0.00}$	$-11.36_{\pm 0.00}$
UniGraph2	$78.14_{\pm 0.45}$	$72.26_{\pm 1.47}$	$83.08_{\pm 0.74}$	$-10.82_{\pm 1.86}$
GIT	$78.33_{\pm 0.33}$	$72.21_{\pm 0.18}$	$83.24_{\pm 0.52}$	$-11.03_{\pm 0.69}$
GFT	$78.16_{\pm 0.01}$	$72.13_{\pm 0.00}$	$83.49_{\pm 0.04}$	$-11.36_{\pm 0.04}$
OFA	$78.60_{\pm 0.57}$	$77.22_{\pm 0.24}$	$79.85_{\pm 0.50}$	**−2.63**$_{\pm 0.61}$
LLaGA	$78.16_{\pm 0.39}$	$72.62_{\pm 0.77}$	$83.84_{\pm 0.27}$	$-11.21_{\pm 0.73}$

G.5 INTERPRETATION — FULL PER-DATASET RESULTS

The main-text interpretation section reports node-level subgraph-ablation attribution fidelity at top-k=10% edges. Here we provide the complementary details: a sparsity sweep over multiple top-k levels, a per-dataset decomposition of Fid$^+$ / Fid$^-$ / characterization score at top-10%, and atom-level graph-classification results on molecular datasets.

Table 16: **Subgraph-ablation sparsity sweep** ($\Delta_{\text{char}} \times 100$, higher is better). $\Delta_{\text{char}} = \text{char}_{\text{grad}} - \text{char}_{\text{rand}}$ at four edge sparsity levels in each test node's K=2-hop receptive field. Cells: mean$_{\pm\text{std}}$ across runs; std for Δ is propagated as $\sqrt{\sigma^2_{\text{grad}} + \sigma^2_{\text{rand}}}$. **Bold** / underline = best / second per dataset and sparsity.

Cora

Method	k=5%	k=10%	k=20%	k=50%
GCN	+9.7$_{\pm 1.2}$	+11.6$_{\pm 2.4}$	+13.0$_{\pm 2.3}$	+9.5$_{\pm 3.2}$
GAT	+9.0$_{\pm 2.2}$	+11.2$_{\pm 2.0}$	+14.7$_{\pm 1.5}$	+13.0$_{\pm 3.9}$
SAGE	+18.9$_{\pm 5.4}$	+30.2$_{\pm 1.7}$	+39.1$_{\pm 3.5}$	+40.0$_{\pm 4.7}$
BGRL	+12.3$_{\pm 1.6}$	+14.2$_{\pm 1.7}$	+16.0$_{\pm 2.6}$	+11.6$_{\pm 3.2}$
GraphMAE	+9.0$_{\pm 1.9}$	+12.8$_{\pm 0.6}$	+15.7$_{\pm 1.2}$	+12.5$_{\pm 2.5}$
UniGraph2	+4.3$_{\pm 4.0}$	+5.4$_{\pm 2.5}$	+5.4$_{\pm 2.3}$	+8.1$_{\pm 1.4}$
GFT	**+20.3**$_{\pm 5.3}$	**+30.6**$_{\pm 3.5}$	**+40.5**$_{\pm 2.8}$	**+40.9**$_{\pm 4.8}$
GIT	+14.5$_{\pm 13.8}$	+21.3$_{\pm 18.6}$	+27.1$_{\pm 23.6}$	+28.1$_{\pm 28.7}$

CiteSeer

Method	k=5%	k=10%	k=20%	k=50%
GCN	+4.5$_{\pm 5.3}$	+5.4$_{\pm 6.4}$	+5.3$_{\pm 7.1}$	+4.5$_{\pm 9.0}$
GAT	+3.0$_{\pm 1.8}$	+3.4$_{\pm 1.9}$	+4.2$_{\pm 2.5}$	+3.7$_{\pm 2.9}$
SAGE	+8.7$_{\pm 6.9}$	+12.0$_{\pm 8.6}$	+15.0$_{\pm 9.5}$	+17.8$_{\pm 12.1}$
BGRL	+8.0$_{\pm 1.2}$	+8.9$_{\pm 0.7}$	+10.0$_{\pm 1.6}$	+7.3$_{\pm 1.0}$
GraphMAE	+5.1$_{\pm 4.4}$	+6.6$_{\pm 4.5}$	+7.6$_{\pm 5.4}$	+7.5$_{\pm 7.3}$
UniGraph2	+1.0$_{\pm 3.3}$	+2.7$_{\pm 3.4}$	+5.4$_{\pm 3.2}$	+4.4$_{\pm 3.0}$
GFT	**+13.9**$_{\pm 20.5}$	**+19.5**$_{\pm 23.9}$	**+22.9**$_{\pm 26.0}$	**+27.6**$_{\pm 35.2}$
GIT	+10.3$_{\pm 3.7}$	+13.6$_{\pm 4.0}$	+16.8$_{\pm 2.9}$	+19.6$_{\pm 3.0}$

PubMed

Method	k=5%	k=10%	k=20%	k=50%
GCN	+12.8$_{\pm 3.4}$	+12.7$_{\pm 2.9}$	+12.7$_{\pm 2.3}$	+8.3$_{\pm 2.6}$
GAT	+17.9$_{\pm 5.4}$	+18.4$_{\pm 3.6}$	+17.4$_{\pm 2.0}$	+12.2$_{\pm 3.8}$
SAGE	+31.0$_{\pm 3.5}$	+33.7$_{\pm 4.1}$	+34.0$_{\pm 3.7}$	+25.8$_{\pm 3.7}$
BGRL	+11.8$_{\pm 4.9}$	+11.0$_{\pm 4.8}$	+10.4$_{\pm 4.6}$	+6.2$_{\pm 5.4}$
GraphMAE	+17.3$_{\pm 2.0}$	+17.3$_{\pm 1.2}$	+17.2$_{\pm 0.4}$	+12.5$_{\pm 1.6}$
UniGraph2	+7.7$_{\pm 6.4}$	+10.7$_{\pm 7.5}$	+14.3$_{\pm 5.7}$	+12.9$_{\pm 6.5}$
GFT	**+36.4**$_{\pm 4.7}$	**+40.1**$_{\pm 5.2}$	**+40.1**$_{\pm 4.7}$	**+31.7**$_{\pm 5.2}$
GIT	+27.8$_{\pm 5.8}$	+30.1$_{\pm 6.0}$	+30.1$_{\pm 5.8}$	+22.4$_{\pm 6.0}$

WikiCS

Method	k=5%	k=10%	k=20%	k=50%
GCN	+26.1$_{\pm 1.3}$	+23.7$_{\pm 2.2}$	+21.7$_{\pm 2.2}$	+18.6$_{\pm 2.2}$
GAT	+26.8$_{\pm 3.5}$	+25.8$_{\pm 4.6}$	+24.4$_{\pm 4.8}$	+19.7$_{\pm 4.8}$
SAGE	+52.5$_{\pm 1.2}$	+51.8$_{\pm 1.6}$	+49.9$_{\pm 1.8}$	+43.1$_{\pm 4.2}$
BGRL	+23.4$_{\pm 0.4}$	+20.1$_{\pm 2.3}$	+18.8$_{\pm 2.4}$	+16.3$_{\pm 2.4}$
GraphMAE	+25.3$_{\pm 2.7}$	+25.4$_{\pm 2.2}$	+23.2$_{\pm 2.6}$	+16.5$_{\pm 2.2}$
UniGraph2	+22.4$_{\pm 4.5}$	+24.6$_{\pm 4.5}$	+22.3$_{\pm 4.0}$	+20.5$_{\pm 3.6}$
GFT	**+64.8**$_{\pm 1.6}$	**+63.7**$_{\pm 1.8}$	**+62.2**$_{\pm 2.0}$	**+55.7**$_{\pm 3.1}$
GIT	+54.1$_{\pm 2.5}$	+53.5$_{\pm 3.1}$	+51.7$_{\pm 3.0}$	+45.9$_{\pm 4.7}$

ElecComp

Method	k=5%	k=10%	k=20%	k=50%
GCN	+13.3$_{\pm 5.0}$	+16.1$_{\pm 6.1}$	+17.6$_{\pm 6.9}$	+18.9$_{\pm 7.6}$
GAT	+6.8$_{\pm 5.4}$	+8.1$_{\pm 5.4}$	+12.6$_{\pm 5.0}$	+17.8$_{\pm 4.8}$
SAGE	+23.1$_{\pm 7.6}$	+25.7$_{\pm 7.1}$	+30.0$_{\pm 5.6}$	+27.3$_{\pm 7.0}$
BGRL	+14.4$_{\pm 7.0}$	+18.2$_{\pm 7.4}$	+20.4$_{\pm 7.7}$	+21.6$_{\pm 7.2}$
GraphMAE	+7.1$_{\pm 7.3}$	+9.1$_{\pm 6.4}$	+14.7$_{\pm 7.3}$	+18.2$_{\pm 7.2}$
UniGraph2	+4.5$_{\pm 6.7}$	+7.3$_{\pm 6.1}$	+8.5$_{\pm 8.2}$	+10.1$_{\pm 7.8}$
GFT	+23.0$_{\pm 6.7}$	+24.9$_{\pm 5.9}$	+28.6$_{\pm 5.1}$	+26.6$_{\pm 5.2}$
GIT	**+24.3**$_{\pm 6.6}$	**+26.5**$_{\pm 6.3}$	**+30.3**$_{\pm 4.5}$	**+27.7**$_{\pm 5.6}$

ElePhoto

Method	k=5%	k=10%	k=20%	k=50%
GCN	+12.2$_{\pm 4.3}$	+16.5$_{\pm 4.9}$	+18.6$_{\pm 4.6}$	+16.9$_{\pm 4.8}$
GAT	+7.8$_{\pm 3.7}$	+9.6$_{\pm 4.4}$	+15.1$_{\pm 1.8}$	+19.6$_{\pm 3.8}$
SAGE	**+23.1**$_{\pm 6.2}$	+24.8$_{\pm 6.6}$	+24.9$_{\pm 6.1}$	**+24.1**$_{\pm 6.5}$
BGRL	+13.9$_{\pm 3.9}$	+17.0$_{\pm 4.3}$	+18.1$_{\pm 4.6}$	+16.3$_{\pm 5.7}$
GraphMAE	+8.0$_{\pm 3.7}$	+10.6$_{\pm 4.5}$	+14.2$_{\pm 3.3}$	+19.8$_{\pm 3.6}$
UniGraph2	+5.8$_{\pm 1.6}$	+6.4$_{\pm 2.6}$	+10.3$_{\pm 3.6}$	+11.4$_{\pm 3.1}$
GFT	+22.2$_{\pm 4.1}$	+24.2$_{\pm 4.6}$	+24.9$_{\pm 4.5}$	+23.3$_{\pm 5.0}$
GIT	+22.7$_{\pm 5.0}$	**+25.0**$_{\pm 4.8}$	**+25.6**$_{\pm 4.4}$	+23.9$_{\pm 5.4}$

AmazonRatings

Method	k=5%	k=10%	k=20%	k=50%
GCN	+3.4$_{\pm 2.7}$	+3.5$_{\pm 3.6}$	+3.6$_{\pm 4.5}$	+2.9$_{\pm 5.9}$
GAT	+1.9$_{\pm 3.4}$	+2.0$_{\pm 4.0}$	+2.1$_{\pm 6.4}$	+2.2$_{\pm 10.1}$
SAGE	+12.4$_{\pm 5.6}$	+16.0$_{\pm 5.7}$	+20.4$_{\pm 4.8}$	+26.2$_{\pm 6.0}$
BGRL	+6.4$_{\pm 5.3}$	+5.9$_{\pm 5.6}$	+7.8$_{\pm 6.6}$	+6.4$_{\pm 6.7}$
GraphMAE	+5.8$_{\pm 2.9}$	+6.7$_{\pm 3.4}$	+6.6$_{\pm 4.8}$	+7.0$_{\pm 5.9}$
UniGraph2	+2.6$_{\pm 5.5}$	+5.0$_{\pm 3.7}$	+6.2$_{\pm 4.8}$	+8.2$_{\pm 2.4}$
GFT	**+15.8**$_{\pm 2.6}$	**+17.3**$_{\pm 2.3}$	**+24.3**$_{\pm 1.4}$	**+29.8**$_{\pm 3.2}$
GIT	+10.3$_{\pm 5.3}$	+14.2$_{\pm 3.1}$	+19.2$_{\pm 5.2}$	+20.6$_{\pm 6.1}$

Tolokers

Method	k=5%	k=10%	k=20%	k=50%
GCN	+22.4$_{\pm 0.3}$	+22.4$_{\pm 0.3}$	+22.3$_{\pm 0.3}$	+21.1$_{\pm 0.3}$
GAT	+0.2$_{\pm 0.0}$	+0.2$_{\pm 0.0}$	+0.2$_{\pm 0.0}$	+0.1$_{\pm 0.0}$
SAGE	+23.6$_{\pm 0.6}$	+23.6$_{\pm 0.6}$	+23.7$_{\pm 0.6}$	+23.2$_{\pm 0.8}$
BGRL	+8.5$_{\pm 1.7}$	+8.2$_{\pm 1.8}$	+7.5$_{\pm 1.9}$	+3.7$_{\pm 2.0}$
GraphMAE	+1.2$_{\pm 0.8}$	+1.2$_{\pm 0.8}$	+1.2$_{\pm 0.8}$	+1.2$_{\pm 0.8}$
UniGraph2	+11.4$_{\pm 4.6}$	+14.2$_{\pm 3.9}$	+17.2$_{\pm 4.6}$	+16.1$_{\pm 5.7}$
GFT	**+30.2**$_{\pm 1.5}$	**+30.2**$_{\pm 1.5}$	**+30.3**$_{\pm 1.4}$	**+29.8**$_{\pm 1.4}$
GIT	+9.9$_{\pm 1.0}$	+9.8$_{\pm 1.0}$	+9.8$_{\pm 1.1}$	+9.3$_{\pm 1.3}$

Table 17: **Subgraph-ablation Fid$^+$ / Fid$^-$ / char decomposition at $k = 10\%$ edges** ($\times 100$). For each dataset/method, columns report the gradient-saliency value (G) and the random-edge baseline (R) for Fid$^+$ (drop in confidence when salient edges are masked; higher is better) and Fid$^-$ (preservation when non-salient edges are masked; lower is better). char = harmonic mean of Fid$^+$ and $1-$Fid$^-$. Δ = char(G) $-$ char(R), the quantity reported in the main-text table; all cells report mean$_{\pm \text{std}}$ over runs (Δ uses propagated std). **Bold** / underline = best / second per dataset on Δ.

Cora

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$5.4_{\pm 2.4}$	$-0.0_{\pm 0.6}$	$-0.9_{\pm 0.4}$	$5.2_{\pm 2.3}$	$14.8_{\pm 2.2}$	$3.2_{\pm 1.1}$	$+11.6_{\pm 2.4}$
GAT	$11.0_{\pm 3.7}$	$2.1_{\pm 1.1}$	$4.9_{\pm 2.4}$	$12.9_{\pm 3.7}$	$16.2_{\pm 1.8}$	$5.0_{\pm 0.9}$	$+11.2_{\pm 2.0}$
SAGE	$24.8_{\pm 1.2}$	$3.2_{\pm 1.1}$	$1.8_{\pm 1.9}$	$27.6_{\pm 3.4}$	$36.2_{\pm 0.7}$	$5.9_{\pm 1.5}$	$\underline{+30.2}_{\pm 1.7}$
BGRL	$7.8_{\pm 1.5}$	$0.5_{\pm 1.2}$	$-1.6_{\pm 3.3}$	$7.6_{\pm 2.9}$	$18.7_{\pm 1.1}$	$4.5_{\pm 1.4}$	$+14.2_{\pm 1.7}$
GraphMAE	$9.0_{\pm 0.9}$	$0.3_{\pm 0.5}$	$5.1_{\pm 3.2}$	$11.6_{\pm 2.8}$	$15.9_{\pm 0.2}$	$3.1_{\pm 0.6}$	$+12.8_{\pm 0.6}$
UniGraph2	$7.4_{\pm 2.4}$	$3.5_{\pm 2.4}$	$13.2_{\pm 0.5}$	$14.6_{\pm 2.2}$	$12.9_{\pm 1.7}$	$7.5_{\pm 1.8}$	$+5.4_{\pm 2.5}$
GFT	$30.3_{\pm 3.7}$	$5.4_{\pm 1.7}$	$6.3_{\pm 2.4}$	$34.8_{\pm 3.5}$	$38.5_{\pm 3.0}$	$7.8_{\pm 1.8}$	**+30.6**$_{\pm 3.5}$
GIT	$17.2_{\pm 13.1}$	$2.2_{\pm 2.0}$	$0.9_{\pm 2.5}$	$18.5_{\pm 14.5}$	$25.4_{\pm 18.3}$	$4.2_{\pm 3.4}$	$+21.3_{\pm 18.6}$

CiteSeer

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$4.0_{\pm 6.7}$	$1.4_{\pm 2.9}$	$-0.8_{\pm 2.6}$	$2.6_{\pm 4.9}$	$9.5_{\pm 5.7}$	$4.1_{\pm 2.8}$	$+5.4_{\pm 6.4}$
GAT	$0.9_{\pm 0.4}$	$-0.0_{\pm 0.4}$	$0.5_{\pm 0.5}$	$1.5_{\pm 0.9}$	$5.7_{\pm 1.8}$	$2.3_{\pm 0.7}$	$+3.4_{\pm 1.9}$
SAGE	$12.1_{\pm 5.7}$	$4.5_{\pm 2.4}$	$0.5_{\pm 0.4}$	$10.8_{\pm 4.1}$	$19.8_{\pm 7.9}$	$7.8_{\pm 3.4}$	$+12.0_{\pm 8.6}$
BGRL	$4.9_{\pm 0.8}$	$0.2_{\pm 0.1}$	$-2.1_{\pm 2.0}$	$3.5_{\pm 1.1}$	$14.2_{\pm 0.6}$	$5.3_{\pm 0.4}$	$+9.0_{\pm 0.7}$
GraphMAE	$4.9_{\pm 4.4}$	$1.2_{\pm 2.8}$	$1.9_{\pm 1.7}$	$3.9_{\pm 3.3}$	$11.6_{\pm 3.6}$	$5.0_{\pm 2.6}$	$+6.6_{\pm 4.5}$
UniGraph2	$5.3_{\pm 4.9}$	$4.2_{\pm 4.3}$	$6.4_{\pm 3.9}$	$9.2_{\pm 5.9}$	$12.7_{\pm 2.4}$	$9.9_{\pm 2.4}$	$+2.8_{\pm 3.4}$
GFT	$27.4_{\pm 21.1}$	$10.2_{\pm 9.2}$	$2.6_{\pm 2.7}$	$23.0_{\pm 16.3}$	$32.6_{\pm 21.7}$	$13.0_{\pm 10.0}$	**+19.5**$_{\pm 23.9}$
GIT	$13.7_{\pm 2.6}$	$4.9_{\pm 1.2}$	$0.5_{\pm 0.2}$	$12.4_{\pm 0.7}$	$22.2_{\pm 3.8}$	$8.6_{\pm 1.4}$	$\underline{+13.6}_{\pm 4.0}$

PubMed

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$8.2_{\pm 3.4}$	$0.9_{\pm 1.0}$	$0.2_{\pm 2.0}$	$9.5_{\pm 1.8}$	$15.1_{\pm 2.7}$	$2.4_{\pm 1.0}$	$+12.7_{\pm 2.9}$
GAT	$18.8_{\pm 2.8}$	$1.9_{\pm 1.4}$	$3.9_{\pm 0.9}$	$17.0_{\pm 0.8}$	$21.6_{\pm 3.3}$	$3.2_{\pm 1.5}$	$+18.4_{\pm 3.6}$
SAGE	$25.3_{\pm 2.6}$	$1.9_{\pm 1.0}$	$1.6_{\pm 0.8}$	$20.7_{\pm 2.4}$	$37.2_{\pm 4.0}$	$3.4_{\pm 1.0}$	$\underline{+33.7}_{\pm 4.1}$
BGRL	$3.8_{\pm 4.7}$	$-0.0_{\pm 0.7}$	$0.2_{\pm 0.9}$	$2.7_{\pm 3.7}$	$13.2_{\pm 4.8}$	$2.1_{\pm 0.8}$	$+11.0_{\pm 4.8}$
GraphMAE	$13.1_{\pm 2.0}$	$0.5_{\pm 0.8}$	$3.9_{\pm 0.4}$	$11.5_{\pm 0.9}$	$19.4_{\pm 0.9}$	$2.1_{\pm 0.7}$	$+17.3_{\pm 1.2}$
UniGraph2	$12.3_{\pm 7.1}$	$0.4_{\pm 2.0}$	$8.5_{\pm 3.1}$	$21.1_{\pm 2.4}$	$15.5_{\pm 7.3}$	$4.7_{\pm 1.8}$	$+10.7_{\pm 7.5}$
GFT	$31.9_{\pm 3.2}$	$1.6_{\pm 1.0}$	$1.4_{\pm 1.5}$	$23.5_{\pm 0.4}$	$43.2_{\pm 5.0}$	$3.1_{\pm 1.3}$	**+40.1**$_{\pm 5.2}$
GIT	$23.3_{\pm 4.0}$	$1.8_{\pm 0.6}$	$1.0_{\pm 0.9}$	$19.2_{\pm 2.4}$	$33.0_{\pm 6.0}$	$2.9_{\pm 0.2}$	$+30.1_{\pm 6.0}$

WikiCS

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$16.2_{\pm 2.6}$	$0.3_{\pm 0.3}$	$-2.7_{\pm 0.3}$	$7.3_{\pm 3.5}$	$26.2_{\pm 2.1}$	$2.5_{\pm 0.8}$	$+23.7_{\pm 2.2}$
GAT	$18.4_{\pm 5.7}$	$-0.1_{\pm 0.5}$	$-0.1_{\pm 1.0}$	$14.3_{\pm 4.7}$	$28.0_{\pm 4.5}$	$2.2_{\pm 0.4}$	$+25.8_{\pm 4.6}$
SAGE	$41.8_{\pm 1.5}$	$0.7_{\pm 0.7}$	$0.1_{\pm 0.1}$	$20.3_{\pm 2.6}$	$54.2_{\pm 1.6}$	$2.4_{\pm 0.5}$	$+51.8_{\pm 1.6}$
BGRL	$11.0_{\pm 3.8}$	$0.4_{\pm 0.4}$	$-0.9_{\pm 0.4}$	$6.4_{\pm 1.9}$	$22.8_{\pm 2.3}$	$2.6_{\pm 0.6}$	$+20.1_{\pm 2.3}$
GraphMAE	$19.1_{\pm 2.6}$	$0.9_{\pm 1.3}$	$-0.1_{\pm 0.2}$	$10.6_{\pm 3.2}$	$28.1_{\pm 1.8}$	$2.7_{\pm 1.3}$	$+25.4_{\pm 2.2}$
UniGraph2	$26.6_{\pm 4.2}$	$3.1_{\pm 0.7}$	$2.9_{\pm 2.5}$	$18.6_{\pm 2.8}$	$31.2_{\pm 4.4}$	$6.6_{\pm 1.0}$	$+24.6_{\pm 4.5}$
GFT	$58.3_{\pm 1.7}$	$1.6_{\pm 0.7}$	$0.1_{\pm 0.2}$	$24.6_{\pm 2.2}$	$66.8_{\pm 1.5}$	$3.2_{\pm 0.9}$	**+63.7**$_{\pm 1.8}$
GIT	$42.9_{\pm 2.4}$	$0.6_{\pm 0.8}$	$0.0_{\pm 0.1}$	$20.7_{\pm 3.2}$	$55.9_{\pm 2.9}$	$2.4_{\pm 0.9}$	$\underline{+53.5}_{\pm 3.1}$

ElecComp

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$18.1_{\pm 5.2}$	$5.7_{\pm 2.7}$	$3.2_{\pm 1.5}$	$18.8_{\pm 6.8}$	$26.0_{\pm 5.2}$	$9.9_{\pm 3.0}$	$+16.1_{\pm 6.1}$
GAT	$10.8_{\pm 1.6}$	$5.9_{\pm 3.3}$	$11.4_{\pm 4.1}$	$20.9_{\pm 0.9}$	$16.8_{\pm 3.2}$	$8.7_{\pm 4.3}$	$+8.1_{\pm 5.4}$
SAGE	$27.9_{\pm 4.2}$	$7.3_{\pm 4.1}$	$5.2_{\pm 2.3}$	$23.4_{\pm 1.7}$	$36.4_{\pm 5.5}$	$10.7_{\pm 4.5}$	$\underline{+25.7}_{\pm 7.1}$
BGRL	$20.2_{\pm 4.9}$	$6.3_{\pm 3.9}$	$4.5_{\pm 1.2}$	$20.5_{\pm 4.7}$	$29.0_{\pm 5.8}$	$10.8_{\pm 4.5}$	$+18.2_{\pm 7.4}$
GraphMAE	$12.5_{\pm 4.6}$	$6.0_{\pm 2.9}$	$14.1_{\pm 4.8}$	$25.4_{\pm 2.1}$	$17.6_{\pm 5.4}$	$8.5_{\pm 3.5}$	$+9.1_{\pm 6.4}$
UniGraph2	$12.5_{\pm 3.8}$	$5.1_{\pm 3.8}$	$16.8_{\pm 2.4}$	$27.7_{\pm 1.6}$	$16.9_{\pm 4.4}$	$9.6_{\pm 4.2}$	$+7.3_{\pm 6.1}$
GFT	$28.1_{\pm 3.4}$	$7.7_{\pm 3.2}$	$4.4_{\pm 1.9}$	$22.7_{\pm 0.5}$	$35.6_{\pm 4.5}$	$10.7_{\pm 3.9}$	$+24.9_{\pm 5.9}$
GIT	$28.5_{\pm 3.7}$	$7.3_{\pm 3.7}$	$4.1_{\pm 1.6}$	$23.2_{\pm 1.4}$	$37.1_{\pm 4.8}$	$10.6_{\pm 4.0}$	**+26.5**$_{\pm 6.3}$

ElePhoto

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$16.2_{\pm 3.7}$	$5.2_{\pm 3.3}$	$4.0_{\pm 0.9}$	$14.5_{\pm 2.2}$	$25.2_{\pm 2.6}$	$8.7_{\pm 4.2}$	$+16.5_{\pm 4.9}$
GAT	$11.0_{\pm 4.0}$	$5.2_{\pm 2.3}$	$8.5_{\pm 3.2}$	$15.6_{\pm 3.0}$	$17.6_{\pm 3.6}$	$8.0_{\pm 2.5}$	$+9.6_{\pm 4.4}$
SAGE	$23.3_{\pm 3.8}$	$4.5_{\pm 4.0}$	$2.1_{\pm 0.9}$	$14.0_{\pm 0.2}$	$32.1_{\pm 4.4}$	$7.2_{\pm 4.9}$	$\underline{+24.8}_{\pm 6.6}$
BGRL	$16.2_{\pm 2.1}$	$5.1_{\pm 2.5}$	$5.4_{\pm 2.4}$	$15.9_{\pm 0.7}$	$25.4_{\pm 2.1}$	$8.4_{\pm 3.7}$	$+17.0_{\pm 4.3}$
GraphMAE	$13.1_{\pm 2.8}$	$6.0_{\pm 3.9}$	$10.0_{\pm 3.7}$	$18.4_{\pm 0.8}$	$18.8_{\pm 2.4}$	$8.2_{\pm 3.8}$	$+10.6_{\pm 4.5}$
UniGraph2	$9.5_{\pm 2.2}$	$3.8_{\pm 2.4}$	$15.6_{\pm 3.5}$	$21.4_{\pm 2.9}$	$14.8_{\pm 2.1}$	$8.4_{\pm 1.5}$	$+6.4_{\pm 2.6}$
GFT	$22.5_{\pm 1.2}$	$4.5_{\pm 3.1}$	$2.5_{\pm 0.6}$	$13.7_{\pm 1.4}$	$31.0_{\pm 1.2}$	$6.8_{\pm 4.4}$	$+24.2_{\pm 4.6}$
GIT	$22.4_{\pm 2.7}$	$4.0_{\pm 3.1}$	$2.1_{\pm 0.7}$	$14.1_{\pm 2.4}$	$31.6_{\pm 2.4}$	$6.6_{\pm 4.1}$	**+25.0**$_{\pm 4.8}$

AmazonRatings

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$3.9_{\pm 3.4}$	$3.5_{\pm 2.1}$	$17.7_{\pm 4.3}$	$15.0_{\pm 4.0}$	$12.0_{\pm 2.6}$	$8.5_{\pm 2.5}$	$+3.5_{\pm 3.6}$
GAT	$1.9_{\pm 2.3}$	$1.1_{\pm 1.0}$	$5.3_{\pm 6.4}$	$4.8_{\pm 5.7}$	$5.6_{\pm 3.7}$	$3.6_{\pm 1.6}$	$+2.0_{\pm 4.0}$
SAGE	$18.6_{\pm 5.2}$	$6.3_{\pm 1.3}$	$15.2_{\pm 1.5}$	$29.2_{\pm 6.9}$	$25.8_{\pm 5.5}$	$9.8_{\pm 1.6}$	$\underline{+16.0}_{\pm 5.7}$
BGRL	$8.4_{\pm 2.2}$	$4.9_{\pm 3.3}$	$22.8_{\pm 0.8}$	$22.7_{\pm 5.0}$	$17.3_{\pm 3.1}$	$11.4_{\pm 4.6}$	$+5.9_{\pm 5.6}$
GraphMAE	$9.2_{\pm 2.8}$	$4.2_{\pm 2.2}$	$15.7_{\pm 0.9}$	$17.1_{\pm 2.0}$	$16.4_{\pm 2.7}$	$9.7_{\pm 2.2}$	$+6.7_{\pm 3.4}$
UniGraph2	$8.1_{\pm 1.6}$	$4.6_{\pm 4.2}$	$22.7_{\pm 5.3}$	$24.4_{\pm 7.1}$	$17.3_{\pm 1.5}$	$12.2_{\pm 3.4}$	$+5.0_{\pm 3.7}$
GFT	$22.0_{\pm 0.9}$	$6.7_{\pm 2.7}$	$16.4_{\pm 1.7}$	$37.1_{\pm 2.1}$	$26.6_{\pm 0.9}$	$9.3_{\pm 2.1}$	**+17.3**$_{\pm 2.3}$
GIT	$16.9_{\pm 2.0}$	$6.3_{\pm 1.6}$	$14.0_{\pm 2.6}$	$27.4_{\pm 1.6}$	$24.2_{\pm 2.8}$	$10.0_{\pm 1.2}$	$+14.2_{\pm 3.1}$

Tolokers

Method	Fid$^+$ G	Fid$^+$ R	Fid$^-$ G	Fid$^-$ R	char G	char R	Δ
GCN	$11.5_{\pm 0.2}$	$0.1_{\pm 0.1}$	$3.0_{\pm 0.6}$	$4.5_{\pm 0.2}$	$22.9_{\pm 0.2}$	$0.4_{\pm 0.2}$	$+22.4_{\pm 0.3}$
GAT	$0.1_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.2_{\pm 0.0}$	$0.0_{\pm 0.0}$	$+0.2_{\pm 0.0}$
SAGE	$13.6_{\pm 0.3}$	$0.2_{\pm 0.2}$	$0.0_{\pm 0.0}$	$3.3_{\pm 0.2}$	$23.9_{\pm 0.5}$	$0.3_{\pm 0.4}$	$\underline{+23.6}_{\pm 0.6}$
BGRL	$-4.1_{\pm 3.9}$	$-0.1_{\pm 0.1}$	$-2.3_{\pm 0.9}$	$1.1_{\pm 3.5}$	$9.5_{\pm 1.7}$	$1.4_{\pm 0.4}$	$+8.2_{\pm 1.8}$
GraphMAE	$0.3_{\pm 0.5}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.1}$	$0.1_{\pm 0.2}$	$1.2_{\pm 0.8}$	$0.0_{\pm 0.0}$	$+1.2_{\pm 0.8}$
UniGraph2	$11.0_{\pm 2.8}$	$0.9_{\pm 2.2}$	$12.4_{\pm 0.6}$	$13.7_{\pm 4.3}$	$20.1_{\pm 3.6}$	$5.9_{\pm 1.7}$	$+14.2_{\pm 3.9}$
GFT	$18.1_{\pm 1.0}$	$0.2_{\pm 0.3}$	$0.0_{\pm 0.0}$	$4.3_{\pm 0.3}$	$30.6_{\pm 1.4}$	$0.4_{\pm 0.5}$	**+30.2**$_{\pm 1.5}$
GIT	$4.5_{\pm 2.0}$	$0.1_{\pm 0.1}$	$0.1_{\pm 0.1}$	$0.4_{\pm 1.6}$	$10.1_{\pm 1.0}$	$0.3_{\pm 0.2}$	$+9.9_{\pm 1.0}$

Table 18: **Graph-level interpretation: $\Delta_{\text{char}} \times 100$ at top-k=20% atom removal.** $\Delta_{\text{char}} = \text{char}_{\text{sal}} - \text{char}_{\text{rand}}$, where char is the GraphFramEx characterization score (Amara et al., 2022) computed from gradient saliency on the fine-tuned model; salient atoms are removed together with their incident edges and compared with random removal of the same number of atoms, so that both conditions undergo the same amount of perturbation. Scores are computed on the positive test graphs that each model predicts correctly. Higher is better. Mean±std over 5 seeds. "—": not applicable to the method's adaptation interface. **Bold** / underline = best / second per column.

Method	Tox21	ChemHIV
GCN	**59.9**$_{\pm 7.2}$	2.6$_{\pm 1.0}$
GAT	48.5$_{\pm 8.6}$	9.7$_{\pm 9.0}$
GraphSAGE	<u>57.0</u>$_{\pm 11.5}$	2.4$_{\pm 2.1}$
BGRL	37.4$_{\pm 6.9}$	**33.9**$_{\pm 5.7}$
GraphMAE	41.0$_{\pm 9.7}$	<u>33.7</u>$_{\pm 3.3}$
UniGraph2	30.9$_{\pm 8.0}$	2.7$_{\pm 1.1}$
GFT	45.4$_{\pm 16.9}$	16.1$_{\pm 19.2}$
GIT	54.7$_{\pm 6.9}$	4.9$_{\pm 1.7}$

H BROADER IMPACTS

GRL-Reliability is intended to inform model selection and risk assessment when graph representation methods are deployed in domains such as citation triage, recommendation, fraud detection, and molecular screening. By making per-axis reliability failures visible alongside clean accuracy, the benchmark can help practitioners avoid overestimating robustness from leaderboard accuracy alone, and can surface capability gaps that motivate new robustness, adaptation, and training objectives.

We do not anticipate direct negative societal impacts. The benchmark relies on publicly available academic datasets, the stress operators are standard noise and distribution-shift constructions rather than novel attacks, and no new pretrained model that meaningfully expands capabilities is released. A residual risk is that practitioners over-interpret aggregate rankings as evidence that any single method is uniformly reliable; the paper emphasizes throughout that reliability behavior is condition-dependent and should be assessed on the axes relevant to a given deployment.