WHEN DO HETEROGENEOUS TEXT-ATTRIBUTED GRAPHS NEED GRAPHS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous text-attributed graphs (HTAGs) combine rich node text with typed graph structure. However, existing HTAG evaluations typically fix a single text encoder, leaving unclear how representation changes affect different structural gains. We introduce **MeHTAGs Bench**, which pairs six HTAG datasets with 15 node representations, including a random control, yielding 90 settings with fixed graphs, labels, target nodes, and splits. Across representation and architecture sweeps, homogeneous message-passing gains generally contract as text representations strengthen, whereas additional relation-aware gains show no consistent contraction. Moreover, the best text-only representation need not be the best graph-model input. We further introduce **Text–Graph Alignment (TGA)**, a label-free, training-free rank statistic that compares the cosine similarities of meta-path-connected target-node pairs with those of randomly sampled pairs. TGA exhibits the strongest absolute rank correlation with residual homogeneous message-passing gains among the evaluated label-free statistics in each of the six HTAGs. Across all six HTAGs, fixed-weight aggregation probes further associate higher input TGA with smaller aggregation-induced alignment gains, even after ceiling normalization. These results show that graph utility depends on the graph–representation pair, with TGA providing an empirical diagnostic of residual homogeneous gains within a fixed HTAG.

1 INTRODUCTION

Many real-world systems combine rich textual descriptions with typed relational structure. Academic networks connect papers, authors, and research fields. Movie platforms organize films, actors, and directors, while patent databases link inventions, inventors, and examiners. Such systems form heterogeneous text-attributed graphs (HTAGs), bringing together textual semantics and heterogeneous connectivity. As text representations continue to improve (Ni et al., 2022; Lee et al., 2025; Zhang et al., 2025), understanding their interaction with graph structure becomes increasingly important.

However, recent HTAG benchmarks (Liu et al., 2025; Zhou et al., 2026; Zhang et al., 2026) typically construct node features using a single text encoder and compare graph architectures within that representation space. Such evaluations establish architectural performance under a particular representation, but leave unresolved how structural gains change when the representation changes. Stronger text representations might intuitively reduce the benefit of graph modeling, yet this intuition does not determine whether gains from homogeneous aggregation and relation-aware aggregation change in the same way, or whether a better text-only representation is also a better graph-model input. We study these distinctions under controlled representation changes and ask whether homogeneous message-passing gains can be measured without labels or downstream predictor training.

To answer these questions, we introduce **MeHTAGs Bench**, a benchmark for **M**ulti-representation **e**valuation of **H**eterogeneous **T**ext-**A**ttributed **G**raph**s**. It pairs six HTAG datasets with 15 node representations, including a random control, static embeddings, pretrained language models, retrieval-oriented encoders, and LLM-based embedding models. Holding the graph, labels, target nodes, and splits fixed yields 90 controlled dataset–representation settings. A representation sweep separates homogeneous gains from additional relation-aware gains through MLP, SAGE, and RSAGE comparisons, while a broader architecture sweep examines the robustness of the observed patterns.

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.

Our experiments show that structural utility is representation-conditioned and selective. Homogeneous gains generally decrease as text-only performance improves, while additional relation-aware gains show no consistent trend across datasets. Better text-only representations also need not yield better graph models. On DBLP, for example, scaling Qwen3 improves MLP performance but reduces RSAGE performance. These findings motivate a diagnostic of residual homogeneous gains that accounts for both representations and graph structure. Classical homophily measures and label informativeness remain fixed when the encoder changes, preventing them from distinguishing representation-dependent gains within an HTAG (Zhu et al., 2020; Platonov et al., 2023). Embedding geometry statistics vary across representations but do not incorporate graph connectivity (Ethayarajh, 2019; Rudman et al., 2022; Garrido et al., 2023).

We introduce **Text–Graph Alignment (TGA)**, a label-free, training-free rank statistic that compares cosine similarities of meta-path-connected target-node pairs with those of randomly sampled pairs. Compared with the evaluated geometry, distinguishability, feature-homophily, and similarity-alignment statistics, TGA achieves the highest mean absolute within-HTAG Spearman correlation with homogeneous gains (0.800 across six HTAGs). Fixed-weight aggregation probes on all six HTAGs further show that higher input TGA is associated with smaller alignment gains from aggregation. This association persists after normalizing for the remaining alignment headroom.

Our main contributions are as follows:

- We introduce **MeHTAGs Bench**, a controlled multi-representation benchmark for evaluating how node representations affect HTAG model performance.
- We show that structural utility is **representation-conditioned and selective**. Homogeneous gains generally contract as text-only performance improves, while additional relation-aware gains show no common trend. Better text-only representations also need not yield better graph models.
- We propose **Text–Graph Alignment (TGA)**, a label-free, training-free rank statistic of representation–structure alignment. TGA achieves the strongest observed absolute Spearman correlation with homogeneous gains among the evaluated label-free statistics in each of the six HTAGs.

2 Related Work

Heterogeneous graph learning. Heterogeneous graph learning methods capture typed structure through relation-specific message passing (Schlichtkrull et al., 2018) and meta-path-based representations (Dong et al., 2017; Wang et al., 2019; Fu et al., 2020). Attention mechanisms and transformer architectures support type-aware aggregation (Hong et al., 2020; Hu et al., 2020), while learned meta-path construction provides an alternative to predefined paths (Yun et al., 2019). Other work systematized heterogeneous representation learning and improved architectural simplicity, scalability, and long-range modeling (Yang et al., 2020; Lv et al., 2021; Yang et al., 2023b; Li et al., 2024a). Yet conventional heterogeneous graph benchmarks largely rely on fixed categorical, bag-of-words, or other shallow features (Zhou et al., 2026), limiting evaluation of how structural utility changes with richer node representations. Building on recent efforts to incorporate raw text into heterogeneous graphs, we examine whether the utility of heterogeneous structure remains stable across representations.

Text-attributed graph learning. Text-attributed graphs (TAGs) attach raw text to nodes and require the joint use of graph structure and textual semantics. Early approaches converted node text into shallow semantic features before applying GNNs. Later work integrated pretrained language models (PLMs) through graph-supervised feature extraction (Chien et al., 2022), joint text–graph transformers (Yang et al., 2021), alternating LM–GNN training (Zhao et al., 2023), and structure-aware pretraining (Jin et al., 2023), with LLMs recently serving as feature enhancers (He et al., 2024). CS-TAG (Yan et al., 2023) systematically compared PLM, GNN, and PLM–GNN co-training paradigms on datasets with original node texts, alongside efforts to improve downstream performance (Wang et al., 2025; Chen et al., 2024b). These studies, however, remain largely confined to the homogeneous setting. Recent HTAG benchmarks cover diverse node and edge types (Liu et al., 2025; Zhou et al., 2026; Zhang et al., 2026), yet each fixes a single text representation per dataset, treating the encoder as an

implementation detail. Whether model rankings and the utility of heterogeneous structure survive a change of text representation thus remains unexamined.

Feature–structure alignment. In TAG modeling, the representation space induced by the text encoder is part of what the model operates on, not merely an input format (Yan et al., 2023). This space varies substantially across encoders, notably in anisotropy that can confound downstream comparisons (Ethayarajh, 2019; Rudman et al., 2022; Liang et al., 2022; Fihey et al., 2026; Gao et al., 2021). As text representations continue to improve (He et al., 2024; Enevoldsen et al., 2025), the boundary between topological information and what the semantic space already encodes remains unclear (Zhang et al., 2021; Li et al., 2024b; Xu et al., 2025). Label-based homophily measures help characterize when graph structure supports node classification (Ma et al., 2021; Luan et al., 2022; 2023), but do not directly assess how typed connectivity aligns with text-derived representations. This gap is particularly sharp in HTAGs, where meta-paths encode semantic structure (Sun et al., 2011; 2022; Liu et al., 2026), whereas meta-path homophily analyses remain on the label side (Guo et al., 2023; Lin et al., 2025). We examine how this alignment varies across text encoders within a fixed HTAG. We quantify it using a label-free, training-free rank statistic and evaluate its association with residual homogeneous message-passing gains.

3 Motivation

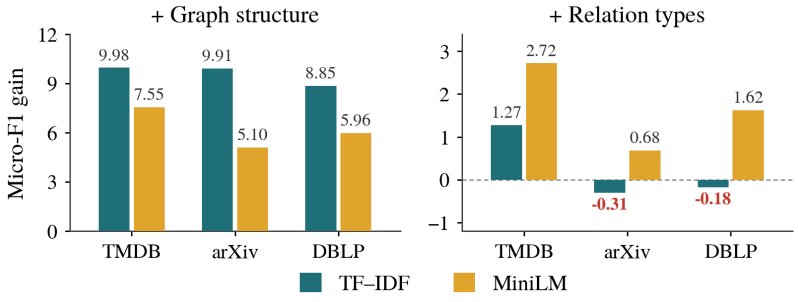


Figure 1: **Structural gains change with the node representation.** Micro-F1 gains on three HTAGs with **TF–IDF** and **MiniLM** node features. Left: gain from adding graph structure (homogeneous SAGE minus MLP). Right: additional gain from incorporating relation types (relation-aware RSAGE minus SAGE).

Most HTAG benchmarks (Liu et al., 2025; Zhou et al., 2026; Zhang et al., 2026) evaluate graph models on features from a single text encoder, implicitly treating graph utility as a fixed property of the dataset. Figure 1 tests this premise through a controlled representation change on TMDB, arXiv, and DBLP. We keep the graph, labels, target nodes, and data splits fixed and replace TF–IDF features with PLM-based MiniLM embeddings. The gain from homogeneous propagation shrinks on all three datasets, while the gain from relation-aware modeling increases on all three and turns from negative to positive on arXiv and DBLP. On arXiv, for example, the homogeneous gain roughly halves, from $+9.91$ to $+5.10$ Micro-F1 points. Structural utility therefore depends on the node representation, with gains from graph structure and relation types moving in opposite directions in this comparison. This motivates evaluating these gains separately across text representations.

4 MeHTAGs Bench

We build MeHTAGs Bench on six HTAG datasets introduced by Liu et al. (2025), spanning movie, community question answering, scholarly, literature, and patent domains across diverse graph scales. Each dataset provides raw node text, and we use fixed time-based splits to evaluate text-based and graph-based models. We retain the original benchmark's 384-dimensional MiniLM features as the provided representation, denoted as Orig. (MiniLM), and add 14 alternative representations. The resulting 15 node representations comprise one random control and 14 text-derived representations, organized into four families: control and static baselines, pretrained language model encoders,

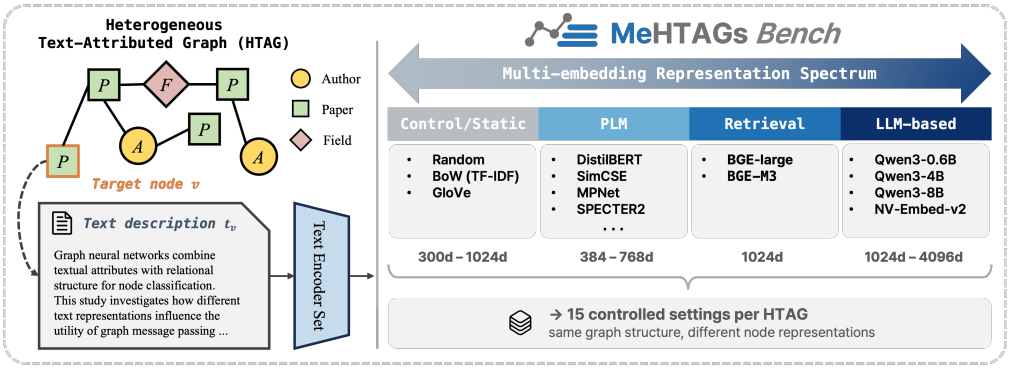


Figure 2: **Overview of MeHTAGs Bench.** Each HTAG is paired with 15 node representations from four representation families, while the graph, labels, target nodes, and data splits remain fixed.

Table 1: **Statistics and representation coverage of MeHTAGs Bench.** Left: graph size, schema, and task of the six HTAGs. Right: number and dimensionality of node representations in each family.

Datasets									Node Representations		
Dataset	Domain	#Nodes	#Edges	#Node Types	#Edge Types	#Targets	#Classes	Task	Family	Count	Dimension
TMDB	Movie	24,412	104,858	3	4	7,505	4	Single-label	Control/Static	3	300–1024
CroVal	CQA	44,386	164,981	3	6	34,153	6	Single-label	PLM	6	384–768
arXiv	Academic	231,111	2,075,692	3	6	81,634	40	Single-label	Retrieval	2	1024
Book	Literature	786,257	9,035,291	3	6	594,484	8	Multi-label	LLM-based	4	1024–4096
DBLP	Academic	1,989,010	29,830,033	3	6	964,350	9	Single-label	**Total**	**15**	300–4096
Patent	Patent	5,646,139	8,833,738	3	4	2,762,187	120	Single-label	**Settings**	**90**	

Notes. Target nodes are movies in TMDB, questions in CroVal, papers in arXiv and DBLP, books in Book, and patents in Patent. Edge counts include both directions of target–target relations and one self-loop per target node, while unique cross-type edges are counted once.

retrieval-oriented encoders, and LLM-based embedding models. As illustrated in Figure 2, we hold the graph, labels, target nodes, and time-based splits fixed while varying the node representation, yielding 90 dataset–representation settings. Table 1 summarizes the dataset statistics and representation coverage, with full representation specifications in Appendix A.

5 Representation-Conditioned Structural Utility

5.1 Preliminaries

Let an HTAG be $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$, where $\mathcal{A}$ and $\mathcal{R}$ denote the sets of node and relation types, and $\phi : \mathcal{V} \to \mathcal{A}$ and $\psi : \mathcal{E} \to \mathcal{R}$ assign node and relation types. Each target node i has raw text t_i and is assigned a representation $x_i^{(e)} \in \mathbb{R}^{d_e}$ under scheme e. For text-derived schemes, $x_i^{(e)} = f_e(t_i)$, whereas control schemes are constructed independently of t_i. We write $X^{(e)}$ for the node-representation matrix and $\hat{x}_i = x_i^{(e)} / \|x_i^{(e)}\|_2$ for an L2-normalized target-node representation. Auxiliary-node features are computed by averaging neighboring target-node features and are recomputed for each representation.

A *meta-path* $m = A_1 \xrightarrow{r_1} \cdots \xrightarrow{r_\ell} A_{\ell+1}$ is a type-level sequence over the network schema (Sun et al., 2011; Sun & Han, 2012). We consider target-returning meta-paths with $A_1 = A_{\ell+1} = \mathcal{T}$, the target-node type. Each meta-path induces a set of target-node pairs. For example, the Paper–Author–Paper (PAP) meta-path connects papers sharing an author.

Throughout MeHTAGs Bench, the graph $\mathcal{G}$, labels $\mathcal{Y}$, target nodes, and splits are held fixed while only $X^{(e)}$ varies, so performance differences across representations can be attributed to changes in the node representation under the fixed evaluation protocol.

5.2 Decomposing Structural Utility

Prior work separates message passing from feed-forward computation (Yang et al., 2023a) and shows that GNN knowledge can be distilled into feature-only predictors (Zhang et al., 2021). These analyses, however, treat graph structure as a single source of gain. In HTAGs, a single performance gap between a GNN and an MLP does not reveal which form of heterogeneous structure contributes to the gain. We therefore distinguish models by the structural information available to their node updates, rather than by architectural nesting. Let $h_v^{(0)} = x_v^{(e)}$, let $\mathcal{N}_r(v)$ denote the neighbors of v under relation r, and let $\mathcal{N}(v) = \bigcup_{r\in\mathcal{R}} \mathcal{N}_r(v)$. The three model categories are characterized as

$$\begin{aligned} \text{MLP:} \quad & h_v^{(\ell+1)} = F^{(\ell)}\left(h_v^{(\ell)}\right), \\ \text{HomoGNN:} \quad & h_v^{(\ell+1)} = F^{(\ell)}\left(h_v^{(\ell)}, \mathrm{AGG}_{u\in\mathcal{N}(v)}\, h_u^{(\ell)}\right), \\ \text{RelGNN:} \quad & h_v^{(\ell+1)} = F^{(\ell)}\left(h_v^{(\ell)}, \left\{\mathrm{AGG}_{u\in\mathcal{N}_r(v)}\, h_u^{(\ell)}\right\}_{r\in\mathcal{R}}\right). \end{aligned} \tag{1}$$

Thus, HomoGNN adds an untyped neighborhood summary, whereas RelGNN retains relation-indexed messages, as in relational graph convolution (Schlichtkrull et al., 2018). For the controlled representation sweep, we use MLP, GraphSAGE (denoted SAGE) (Hamilton et al., 2017), and its relation-aware extension HeteroSAGE (denoted RSAGE), following Liu et al. (2025).

For a fixed dataset and representation, we define two contrasts in Micro-F1: $\Delta_{\mathrm{homo}} = \mathrm{Perf}(\mathrm{SAGE}) - \mathrm{Perf}(\mathrm{MLP})$, the observed homogeneous-model gain, and $\Delta_{\mathrm{rel}} = \mathrm{Perf}(\mathrm{RSAGE}) - \mathrm{Perf}(\mathrm{SAGE})$, the observed additional relation-aware-model gain. These contrasts compare model implementations and do not isolate the effects of structural information from other architectural differences. We use these contrasts throughout the representation sweep and the TGA analyses in Section 7. At five representative tiers, a separate architecture sweep tests whether the same qualitative conclusions hold across broader model families.

6 Text–Graph Alignment

6.1 Definition

TGA measures how strongly typed connectivity is reflected in a node representation e within a given HTAG. For a target-to-target meta-path m, we sample anchors and follow its relation sequence by uniform random walks. Successful walks form a connected pair set $\mathcal{P}_m$, while an equally sized reference set $\mathcal{R}_m$ consists of independently sampled distinct target-node pairs. For $(u, v) \in \mathcal{P}_m$ and $(u', v') \in \mathcal{R}_m$, let $s_m^+ = \hat{x}_u^{(e)\top} \hat{x}_v^{(e)}$ and $s_m^- = \hat{x}_{u'}^{(e)\top} \hat{x}_{v'}^{(e)}$. We define

$$\mathrm{TGA}(e, m) = \Pr(s_m^+ > s_m^-) + \tfrac{1}{2}\Pr(s_m^+ = s_m^-), \tag{2}$$

the normalized Mann–Whitney U statistic, equivalently the AUC obtained by using cosine similarity to distinguish meta-path-connected from random pairs. Thus, high TGA indicates that typed connectivity between node pairs is already strongly reflected in representation space. Writing $U(X, Y) = \Pr(X > Y) + \frac{1}{2}\Pr(X = Y)$, we have $\mathrm{TGA}(e, m) = U(s_m^+, s_m^-)$. The symmetry $U(Y, X) = 1 - U(X, Y)$ implies $U(X, X') = \frac{1}{2}$ for i.i.d. X, X', so the tie correction fixes the null at $\frac{1}{2}$ even for discrete or sparse similarities. Because U depends only on pairwise ordering, TGA is invariant under strictly increasing transformations of the similarity score and therefore avoids reliance on encoder-specific absolute cosine levels. For a representation-level score, we macro-average over the complete enumerated set $\mathcal{M}$ of target-returning meta-paths: $\mathrm{TGA}(e) = |\mathcal{M}|^{-1} \sum_{m\in\mathcal{M}} \mathrm{TGA}(e, m)$.

Although TGA is label-free, labels provide a useful null for interpreting its value. Let $h_m = \Pr(y_u = y_v \mid (u, v) \in \mathcal{P}_m)$ denote meta-path label homophily, and let $h_0 = \Pr(y_{u'} = y_{v'} \mid (u', v') \in \mathcal{R}_m)$ denote chance label agreement under the reference-pair sampler. Let $A(e) = U(s^{\mathrm{same}}, s^{\mathrm{diff}})$ denote class separability in representation e. We consider a deliberately restrictive label-only null in which labels fully mediate similarity. Assumption (A1) states that, conditioned on whether a pair shares a class, connected and random pairs draw similarities from the same distribution. Assumption (A2) states that, within same-class and within different-class pairs, the two samplers have the same class

composition. Under these assumptions, s_m^+ and s_m^- are mixtures of the same two distributions, differing only in their same-class weights h_m and h_0, and

$$\mathrm{TGA}(e, m) = \tfrac{1}{2} + \Big(A(e) - \tfrac{1}{2}\Big)\Big(h_m - h_0\Big). \tag{3}$$

equation 3 separates centered TGA into a representation factor $A(e) - \frac{1}{2}$, how well representation e separates classes, and a graph factor $h_m - h_0$, how much more often meta-path m connects same-class nodes than chance. Under this null, TGA equals $\frac{1}{2}$ whenever either factor is zero and exceeds $\frac{1}{2}$ when the two factors have the same sign. For $h_m \geq h_0$, the bound $A(e) \leq 1$ implies $\mathrm{TGA}(e, m) \leq \frac{1}{2} + \frac{1}{2}(h_m - h_0)$. We do not treat this equality as a generative law for real HTAGs. Rather, it defines a label-only null. We compare empirical TGA values with this bound to examine departures from the label-only null and assess the scope of this reference in real HTAGs. The factorization relates alignment to label agreement and class separability under explicit assumptions, while TGA itself remains directly computable from node representations and typed connectivity without requiring those assumptions to hold. Full sampling procedures, meta-path validity criteria, derivations, and empirical analyses are provided in Appendix D.

6.2 TGA Across MeHTAGs Bench

Table 2: **Text–Graph Alignment across MeHTAGs Bench.** Values are $\mathrm{TGA}(e)$ macro-averaged over the enumerated target-returning meta-paths. Bold indicates the highest TGA within each dataset.

	Control	Static		PLM						Retrieval		LLM-based			
Dataset	Random	TF–IDF	GloVe	DistilBERT	SimCSE	MPNet	SPECTER2	MiniLM	Sentence-T5	BGE-large	BGE-M3	Qwen3-0.6B	Qwen3-4B	Qwen3-8B	NV-Embed-v2
TMDB	0.500	0.539	0.556	0.599	0.606	0.603	0.546	0.595	0.614	0.605	0.598	0.611	0.627	0.637	**0.648**
CroVal	0.499	0.714	0.605	0.607	0.653	0.778	0.754	0.767	0.696	0.763	0.727	0.772	**0.790**	0.788	0.786
arXiv	0.500	0.738	0.682	0.703	0.706	**0.826**	0.819	0.818	0.751	0.810	0.759	0.809	0.821	0.824	0.801
Book	0.501	0.630	0.627	0.677	0.699	0.722	0.628	0.700	0.693	0.714	0.707	0.724	0.739	0.742	**0.773**
DBLP	0.499	0.747	0.705	0.741	0.736	**0.851**	0.843	0.843	0.772	0.838	0.777	0.821	0.841	0.843	0.825
Patent	0.501	0.798	0.812	0.846	0.867	0.867	0.874	0.872	0.871	0.878	0.859	0.900	0.902	**0.915**	0.899

Across the six HTAGs, we enumerate 92 target-returning meta-paths (Table 17) and compute TGA for all 15 representations, obtaining 1,380 representation–meta-path measurements. Table 2 reports the resulting representation-level macro-averages. The random control remains near the null value across all six HTAGs (0.499–0.501), whereas text-derived representations vary substantially across graph–representation pairs. TGA characterizes the graph–representation pair. For example, SPECTER2 representations yield TGA values of 0.546 on TMDB, 0.819 on arXiv, and 0.843 on DBLP. Even for a single node representation, TGA quantifies how consistently meta-path-connected pairs are more similar than random pairs, providing an interpretable measure of alignment without labels or predictor training. We test whether higher TGA is associated with smaller residual homogeneous message-passing gains across representations within a fixed HTAG (Section 7).

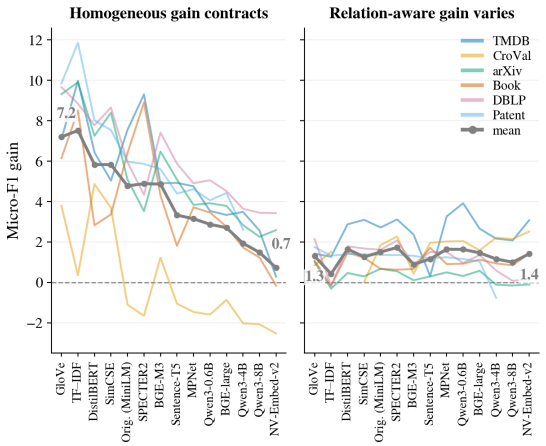
7 Experiments

7.1 Experimental Setup

We evaluate on MeHTAGs Bench (Section 4), comprising six HTAGs paired with 15 representations. The representation sweep runs MLP, SAGE, and RSAGE over all tiers, while the architecture sweep evaluates 11 models at five representative tiers. All runs use the same time-based splits and training protocol, with five seeds. We report Micro-F1 and Macro-F1. Implementation details and complete results are provided in Appendices B and C.

7.2 Representation Strength and Residual Homogeneous Headroom

Figure 3 shows structural gains across node representations ordered by mean within-HTAG text-only MLP rank. Across the five HTAGs with completed results at both endpoints, mean Δ_{homo} contracts sharply from GloVe to NV-Embed-v2, whereas mean Δ_{rel} changes little and shows no common direction across datasets. Within each HTAG, homogeneous headroom is also lower at the strongest completed endpoint than at the weakest. These results show that representation strength is associated more consistently with homogeneous structural utility than with the additional benefit of relation-aware modeling. Let $\Delta_{\text{best}} = \max\{\text{Perf}(\text{SAGE}), \text{Perf}(\text{RSAGE})\} - \text{Perf}(\text{MLP})$ denote the best observed gain over the text-only baseline among the two graph models in this comparison. TGA is negatively associated with both Δ_{homo} and Δ_{best} in all six HTAGs under Pearson and Spearman correlations, while its association with Δ_{rel} has mixed signs. TGA therefore ranks both residual homogeneous headroom and the best observed graph-model gain in this controlled comparison.



TGA and residual graph gains

Dataset	Pearson r			Spearman ρ		
	Δ_{homo}	Δ_{rel}	Δ_{best}	Δ_{homo}	Δ_{rel}	Δ_{best}
TMDB	-0.93^{*}	$+0.21$	-0.87^{*}	-0.87^{*}	$+0.05$	-0.90^{*}
CroVal†	-0.94^{*}	$+0.74^{*}$	-0.95^{*}	-0.89^{*}	$+0.75^{*}$	-0.93^{*}
arXiv	-0.88^{*}	-0.23	-0.88^{*}	-0.78^{*}	-0.07	-0.79^{*}
Book	-0.85^{*}	$+0.39$	-0.89^{*}	-0.71^{*}	$+0.11$	-0.75^{*}
DBLP	-0.91^{*}	-0.23	-0.86^{*}	-0.72^{*}	-0.24	-0.70^{*}
Patent	-0.94^{*}	-0.54	-0.93^{*}	-0.83^{*}	-0.63	-0.85^{*}
Mean	**-0.91**	$+0.06$	**-0.90**	**-0.80**	≈ 0	**-0.82**
Neg./Pos.	6/0	3/3	6/0	6/0	3/3	6/0

$^{*}p < 0.01$, unadjusted two-sided permutation tests (20,000 permutations). $n = 14$ text-derived representations, except Patent ($n = 12$). †CroVal uses $n = 13$ for Δ_{rel} and Δ_{best}.

Figure 3: **Representation-conditioned structural utility.** Structural gains across node representations (left) and within-HTAG Pearson and Spearman correlations between TGA and these gains (right). Mean curves use the five HTAGs with complete endpoint results.

Architecture-sweep robustness. The architecture sweep tests whether the pattern in Figure 3 extends across 11 models spanning text-only, untyped message-passing, typed message-passing, meta-path, and Transformer-based families at five representative representations (Appendix C.2). Among the five HTAGs with complete results for the MLP and the compared untyped and typed models, the mean gain of the best untyped model over the MLP falls from 4.78 to 0.77 Micro-F1 points between Orig. (MiniLM) and NV-Embed-v2. The best typed model's mean advantage over the best untyped model remains nearly unchanged, moving from 1.64 to 1.65 points. These results show that the selective contraction of homogeneous gains persists when a broader set of graph models is considered. After centering both TGA and performance within each HTAG and pooling each model's completed settings, Micro-F1 correlations are positive for all 11 models, ranging from $r = 0.70$ to 0.84, with comparable Macro-F1 results. On arXiv and DBLP, TGA aligns more closely with representation rankings under relation-aware models than under the MLP. These results do not establish that TGA identifies the best architecture or predicts the magnitude of gains over the text-only baseline.

7.3 TGA Compared with Existing Statistics

For a fixed HTAG, exhaustively training graph models for every candidate text representation e is costly. Using text-only MLP performance as a proxy also complicates interpretation, since Δ_{homo} is defined relative to the MLP. We therefore ask whether a statistic computed from embeddings and typed connectivity alone, before predictor training and without labels, can rank Δ_{homo} across candidate representations. We compare TGA with statistics of graph structure, graph–label dependence, embedding geometry, label distinguishability, feature homophily, and similarity alignment. Class separation $A(e)$ and MLP performance serve as label-informed references. For each HTAG, we

compute Spearman correlations with Δ_{homo} and the additional relation-aware gain, Δ_{rel}. Table 3 reports these correlations and their macro-averaged absolute values. We summarize their difference as $\text{Corr. gap} = \overline{|\rho(\cdot, \Delta_{\mathrm{homo}})|} - \overline{|\rho(\cdot, \Delta_{\mathrm{rel}})|}$. A positive gap indicates a stronger average association with homogeneous gain, while a negative gap indicates a stronger average association with additional relation-aware gain.

The structure and graph–label dependence statistics remain fixed as the node representation changes and therefore cannot rank candidate representations within an HTAG. Geometry statistics vary with e but do not incorporate graph connectivity. Among them, RankMe (Garrido et al., 2023) achieves the highest mean absolute correlation with Δ_{homo} (0.596), compared with 0.170 for anisotropy (Ethayarajh, 2019). These statistics characterize the embedding space but do not directly measure how it aligns with graph connectivity. The label-informed references assess how well representations separate or predict labels. Class separation $A(e)$ measures label distinguishability, while MLP performance reflects what the representation alone resolves. Both require labels, and the latter additionally requires training a predictor for each representation. Although text-only performance has the strongest average absolute association with Δ_{homo} in the table, this association requires care because the gain is defined relative to that same baseline. Neither reference directly measures representation–structure alignment.

Across all six HTAGs, higher TGA is associated with smaller residual homogeneous message-passing gains. This association is statistically significant in every HTAG under two-sided permutation tests at the 0.01 level and remains negative in all 24 leave-one-representation-family-out analyses. TGA shows the strongest observed absolute Spearman correlation among the evaluated label-free statistics in every HTAG. Its advantage over the matched typed mean-gap control, which uses the same inputs, supports the value of rank-based alignment over mean similarity differences. Pearson, Spearman, and Kendall results are provided in Appendix H.

Table 3: **Ranking homogeneous headroom with existing statistics and TGA.** Within-HTAG Spearman correlations across text-derived representations, with cross-HTAG summaries.

		Inputs			Spearman ρ with Δ_{homo}						Summary		
Category	Statistic	$X^{(e)}$	Graph	$\mathcal{Y}$	TMDB	CroVal	arXiv	Book	DBLP	Patent	$\overline{\lvert\rho_{\mathrm{h}}\rvert}$	$\overline{\lvert\rho_{\mathrm{r}}\rvert}$	Corr. gap
Structure	homophily & related†	–	✓	✓			constant in e				–	–	–
Graph–label dependence	label informativeness LI	–	✓	✓			constant in e				–	–	–
Geometry	anisotropy	✓	–	–	0.22	0.38	0.13	0.06	0.15	0.06	0.170	0.327	−0.157
	IsoScore	✓	–	–	0.56	0.36	0.68	0.67	0.45	0.01	0.456	0.362	0.094
	RankMe	✓	–	–	−0.71	−0.66	−0.54	−0.54	−0.59	−0.54	0.596	0.540	0.056
Label distinguishability	dist. ratio J	✓	U	✓	0.19	0.46	0.16	0.66	−0.16	−0.31	0.323	**0.621**	−0.298
Feature homophily	Tri-Hom h_E^*‡	✓	U	✓	−0.49	−0.69	−0.67	−0.39	−0.58	−0.39	0.536	0.287	0.248
Similarity alignment	edge sim.	✓	U	–	0.22	0.37	0.17	0.06	0.17	0.06	0.177	0.336	−0.159
	mean gap	✓	U	–	−0.65	−0.64	−0.53	−0.40	−0.57	−0.18	0.496	0.269	0.227
	mean gap	✓	T	–	−0.72	−0.61	−0.55	−0.37	−0.55	−0.19	0.499	0.256	0.242
Rank alignment	**TGA (ours)**	✓	T	–	**−0.87**	**−0.89**	**−0.78**	**−0.71**	**−0.72**	**−0.83**	**0.800**	0.310	**0.490**
Label-informed references	class sep. $A(e)$	✓	–	✓	−0.81	−0.90	−0.70	−0.75	−0.72	−0.70	0.764	0.289	0.474
	MLP	✓	–	✓	−0.93	−0.93	−0.98	−0.98	−0.99	−0.99	0.965	0.510	0.454

Bold highlights TGA's per-HTAG correlations and summary maxima excluding MLP. Summary columns average $|\rho|$ over HTAGs for Δ_{homo} and Δ_{rel}, with Corr. gap equal to the former minus the latter. U: untyped target-node projection. T: typed target-returning meta-path connectivity. The distance ratio is computed on aggregated features, $J(\widetilde{A}_U X^{(e)})$. †Representation-invariant statistics, so within-HTAG correlations are undefined. ‡The Tri-Hom estimator uses labels although its estimand is label-free. Full values are in Appendices F and G.

7.4 What Does TGA Capture?

Aggregation-induced alignment. We hypothesize that, within a fixed HTAG, node representations with higher input TGA receive smaller alignment gains from homogeneous aggregation. To test this hypothesis, we intervene on a trained two-layer SAGE while holding its weights fixed. This allows us to examine how activating aggregation changes alignment within the same trained model without retraining. We compare its final 128-dimensional pre-classifier representations with neighbor aggregation enabled in both layers, Z_{on}, and disabled in both layers, Z_{off}. Both conditions use the same graph-trained weights, so Z_{off} is not a separately trained MLP. We measure the raw

change, $\Delta\mathrm{TGA}_{\mathrm{raw}} = \mathrm{TGA}(Z_{\mathrm{on}}) - \mathrm{TGA}(Z_{\mathrm{off}})$, and its ceiling-normalized form $\Delta\mathrm{TGA}_{\mathrm{norm}} = \Delta\mathrm{TGA}_{\mathrm{raw}}/[1 - \mathrm{TGA}(Z_{\mathrm{off}})]$.

The normalized form expresses the alignment change relative to the remaining distance from $\mathrm{TGA}(Z_{\mathrm{off}})$ to its upper bound of one. Aggregation increases TGA in every completed combination of HTAG and node representation. Table 4 shows that higher input TGA is associated with smaller aggregation-induced alignment gains across all six HTAGs, both before and after ceiling normalization. For both measures, all correlations are significant at the 0.05 level under unadjusted one-sided permutation tests. The persistence after normalization indicates that the pattern is not explained solely by the smaller available range of high-TGA inputs. This model-internal evidence supports the interpretation that already aligned inputs leave less additional alignment for homogeneous aggregation to introduce. How this representational effect translates into predictive gains remains to be established.

Table 4: **Input TGA and alignment gain.**

HTAG	ρ_{raw}	ρ_{norm}
TMDB	-0.846	-0.802
CroVal	-0.789	-0.877
arXiv	-0.710	-0.727
Book	-0.662	-0.714
DBLP	-0.516	-0.552
Patent	-0.734	-0.510

Representation quality, model scale, and dimensionality. We next test whether TGA is primarily driven by output dimensionality. Larger Qwen3 models also produce wider native embeddings, so we separate these factors using Matryoshka representations (Zhang et al., 2025; Kusupati et al., 2022) (Figure 4). Scaling Qwen3 from 0.6B to 8B at a fixed output width of 1,024 dimensions raises TGA and reduces Δ_{homo} on all six HTAGs. Holding Qwen3-8B fixed while varying only its output width changes TGA comparatively little and produces mixed changes in homogeneous headroom. These controls show that, within the evaluated Qwen3 configurations, TGA varies more with encoder scale than with output width. The associated TGA–headroom pattern therefore cannot be explained by output dimensionality alone. Full results appear in Appendix E.

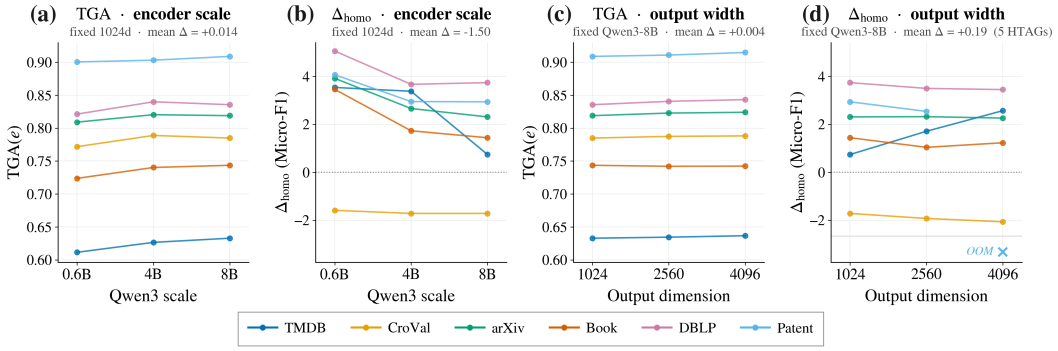


Figure 4: **Dimension-controlled Qwen3 scaling. (a–b)** Varying model size (0.6B–8B) at a fixed 1,024-dimensional output. **(c–d)** Varying output width for Qwen3-8B. TGA tracks model size more than output width. Patent at 4,096 dimensions is omitted from (d) due to OOM.

8 Conclusion

Our study addresses how node representations change the utility of graph structure and relation types in HTAGs. **MeHTAGs Bench** evaluates this dependence across 15 node representations on six HTAGs under fixed graphs, labels, and splits. Stronger representations generally reduce homogeneous headroom, while additional relation-aware gains show no common trend. This distinction makes it necessary to evaluate what graph structure and relation types each contribute as node representations change. Using embeddings and typed connectivity without labels or predictor training, **TGA** provides a stronger within-HTAG ranking signal for residual homogeneous headroom than the evaluated label-free alternatives. It also ranks the best observed gain over the text-only baseline among the graph models in the controlled comparison, supporting evaluation prioritization across candidate node representations. Accordingly, results obtained with a single encoder should be interpreted as properties of that encoder–graph pairing, rather than as general evidence for the utility of the graph's structure or relation types.

AI USE STATEMENT

We used generative AI tools for brainstorming, providing feedback on experimental methodology, assisting with the implementation of parts of the experimental methods, reviewing mathematical formulations and derivations, interpreting experimental results, and drafting and refining portions of the manuscript. All AI-assisted contributions were critically reviewed by the authors, including checks of mathematical arguments, numerical statements, and consistency with the implementation. The authors take full responsibility for the final content of this work, including all claims, experimental results, and artifacts.

ETHICS STATEMENT

This work evaluates graph learning methods using publicly available benchmark datasets. It does not involve human-subject research, collection of personal data, or deployment in high-stakes decision-making settings. We follow the licenses and intended uses of the datasets and acknowledge that benchmark-based evaluations may inherit biases or limitations present in the underlying data.

REPRODUCIBILITY STATEMENT

We plan to release MeHTAGs, including the benchmark data, data-processing pipelines, and the implementation used in our experiments. To support reproducibility without requiring computationally expensive reruns, we will also release the experimental artifacts generated in our study, including the original experimental configurations from Liu et al. (2025) and those for TGA and the baselines. These materials will enable verification and extension of the experimental setup and results reported in this paper.

REFERENCES

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024a.

Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. Label-free node classification on graphs with large language models (llms). In *International Conference on Learning Representations*, volume 2024, pp. 31632–31655, 2024b.

Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S Dhillon. Node feature extraction by self-supervised multi-scale neighborhood prediction. 2022.

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 135–144, 2017.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, et al. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 101715–101771, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Pierre Fihey, Matthieu Labeau, and Pavlo Mozharovskyi. Enhancing two steps textual anomaly detection through anisotropy mitigation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California, United States, July

2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1312. URL https://aclanthology.org/2026.acl-long.1312/.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of the web conference 2020*, pp. 2331–2341, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Jiayan Guo, Lun Du, Wendong Bi, Qiang Fu, Xiaojun Ma, Xu Chen, Shi Han, Dongmei Zhang, and Yan Zhang. Homophily-oriented heterogeneous graph rewiring. In *Proceedings of the ACM web conference 2023*, pp. 511–522, 2023.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *International conference on learning representations*, volume 2024, pp. 5711–5732, 2024.

Huiting Hong, Hantao Guo, Yucheng Lin, Xiaoqing Yang, Zang Li, and Jieping Ye. An attention-based graph neural network for heterogeneous structural learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4132–4139, 2020.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, pp. 2704–2710, 2020.

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. Patton: Language model pretraining on text-rich networks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7005–7020, 2023.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Chao Li, Zijie Guo, Qiuting He, and Kun He. Long-range meta-path search on large-scale heterogeneous graphs. *Advances in Neural Information Processing Systems*, 37:44240–44268, 2024a.

Zhuofeng Li, Zixing Gou, Xiangnan Zhang, Zhongyuan Liu, Sirui Li, Yuntong Hu, Chen Ling, Zheng Zhang, and Liang Zhao. Teg-db: A comprehensive dataset and benchmark of textual-edge graphs. *Advances in Neural Information Processing Systems*, 37:60980–60998, 2024b.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in neural information processing systems*, 34:20887–20902, 2021.

Junhong Lin, Xiaojie Guo, Shuaicheng Zhang, Yada Zhu, and Julian Shun. When heterophily meets heterogeneity: Challenges and a new large-scale graph benchmark. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5607–5618, 2025.

Yunhui Liu, Qizhuo Xie, Jinwei Shi, Jiaxu Shen, and Tieke He. Multi-scale heterogeneous text-attributed graph datasets from diverse domains. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 757–760, 2025.

Yunhui Liu, Yongchao Liu, Yinfeng Chen, Chuntao Hong, Tao Zheng, and Tieke He. Learning hierarchical knowledge in text-rich networks with taxonomy-informed representation learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 961–972, 2026.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. *Advances in neural information processing systems*, 35:1362–1375, 2022.

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. *Advances in Neural Information Processing Systems*, 36:28748–28760, 2023.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 1150–1160, 2021.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the association for computational linguistics: ACL 2022*, pp. 1864–1874, 2022.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 36:523–548, 2023.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. Isoscore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3325–3339, 2022.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL `https://arxiv.org/abs/1910.01108`.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, 2023.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Yizhou Sun and Jiawei Han. *Mining heterogeneous information networks: principles and methodologies*, volume 5. Morgan & Claypool Publishers, 2012.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4 (11):992–1003, 2011.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Heterogeneous information networks: the past, the present, and the future. *Proceedings of the VLDB Endowment*, 15(12), 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Haoyu Peter Wang, Shikun Liu, Rongzhe Wei, and Pan Li. Generalization principles for inference over text-attributed graphs with large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

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

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. Sgformer: Simplifying and empowering transformers for large-graph representations. *Advances in neural information processing systems*, 36:64753–64773, 2023.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Haotian Xu, Yuning You, and Tengfei Ma. When structure doesn't help: Llms do not read text-attributed graphs as effectively as we expected. *arXiv preprint arXiv:2511.16767*, 2025.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, et al. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems*, 36:17238–17264, 2023.

Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4854–4873, 2020.

Chenxiao Yang, Qitian Wu, Jiahua Wang, and Junchi Yan. Graph neural networks are inherently good generalizers: Insights by bridging GNNs and MLPs. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=dqnNW2omZL6.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810, 2021.

Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 10816–10824, 2023b.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

Chenghao Zhang, Qingqing Long, Ludi Wang, Wenjuan Cui, Jianjun Yu, and Yi Du. Cite: Benchmarking heterogeneous text-attributed graph models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31426–31448, 2026.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*, 2021.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

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

Yilun Zheng, Sitao Luan, and Lihui Chen. What is missing for graph homophily? disentangling graph homophily for graph neural networks. *Advances in Neural Information Processing Systems*, 37:68406–68452, 2024.

Lixin Zhou, Zemin Liu, Yuan Fang, Dan Niu, and Jing Ying. Thgb: A comprehensive benchmark for text-attributed heterogeneous graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28973–28981, 2026.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

Appendix

A Node Representations **16**

B Implementation Details **17**

C Full Experimental Results **17**

C.1 Full Representation Sweep 17

C.2 Full Architecture Sweep 21

D Additional Analyses of TGA **29**

D.1 Enumerated Meta-Paths 29

D.2 TGA Computation 29

D.3 Derivation of the Label-Only Null 30

D.4 Empirical Departures from the Label-Only Null 31

D.5 Sampling Stability and Within-HTAG Calibration 31

D.6 Low-Rank Structure of the TGA Landscape 32

E Dimension-Controlled Qwen3 Analysis **33**

F Representation-Invariant Statistics **35**

G Representation-Dependent Statistics **37**

H Correlation Robustness Across Coefficients **41**

I Geometry of Node Representations **42**

J Limitations and Future Work **49**

A NODE REPRESENTATIONS

We construct node representations using a random control, text feature construction methods, and pretrained text encoders. The dimensionalities reported below correspond to the actual node features used in our experiments. All features are ℓ_2-normalized. Transformer inputs are truncated to 512 tokens (384 for MPNet, its native limit), and retrieval-style encoders (BGE-large, Qwen3-Embedding, NV-Embed-v2) encode node texts on the document side without an instruction prefix.

Random. As a control, we assign each node a randomly sampled Gaussian feature vector. We use 384-dimensional vectors.

TF–IDF. TF–IDF encodes lexical term importance using term frequency weighted by inverse document frequency (Salton & Buckley, 1988). We keep the 1024 most frequent terms after English stop-word removal, use sublinear term frequency, and fit the vocabulary on training nodes only.

GloVe. GloVe learns static word vectors from global word–context co-occurrence statistics; we mean-pool word vectors to obtain text features. We use 300-dimensional GloVe embeddings (Pennington et al., 2014).

DistilBERT. DistilBERT compresses BERT through knowledge distillation while retaining contextualized language representations. We mean-pool token representations from `distilbert-base-uncased` to obtain 768-dimensional features (Sanh et al., 2020).

SimCSE. SimCSE learns semantically meaningful sentence representations through contrastive learning. We use `sup-simcse-bert-base-uncased`, which produces 768-dimensional embeddings (Gao et al., 2021).

MPNet. MPNet combines masked and permuted language-modeling objectives to exploit bidirectional context. We use `all-mpnet-base-v2`, which produces 768-dimensional sentence embeddings (Song et al., 2020).

SPECTER2. SPECTER2 is a scientific-document encoder trained with citation-informed supervision (Singh et al., 2023). We use the base model `specter2_base` without task-specific adapters, with CLS pooling, producing 768-dimensional embeddings.

Orig. (MiniLM). MiniLM is a compact Transformer obtained by distilling self-attention relations from a larger teacher model. We use the provided `all-MiniLM-L6-v2` node features, which are included in the graph data and have 384 dimensions (Wang et al., 2020).

Sentence-T5. Sentence-T5 adapts pretrained text-to-text models into sentence encoders through contrastive fine-tuning. We use `sentence-t5-xl`, which produces 768-dimensional embeddings (Ni et al., 2022).

BGE-large. BGE is a retrieval-oriented text encoder trained with contrastive learning and hard negatives. We use `bge-large-en-v1.5` with 1024-dimensional embeddings (Xiao et al., 2024).

BGE-M3. BGE-M3 supports multilingual dense, sparse, and multi-vector retrieval through unified fine-tuning and self-knowledge distillation. We use its dense 1024-dimensional representation (Chen et al., 2024a).

Qwen3-Embedding. Qwen3-Embedding is an instruction-aware LLM-based encoder trained for multilingual and cross-lingual text retrieval (Zhang et al., 2025). We include three scales and use each model's native output width without Matryoshka truncation: 1024 dimensions for Qwen3-Embedding-0.6B, 2560 for Qwen3-Embedding-4B, and 4096 for Qwen3-Embedding-8B.

NV-Embed-v2. NV-Embed-v2 is an LLM-based generalist embedding model trained with contrastive objectives, hard-negative mining, and task instructions (Lee et al., 2025). We use its 4096-dimensional latent-attention-pooled embeddings, encoding node texts without a task instruction for consistency with the other encoders.

B Implementation Details

Models are implemented in PyTorch with DGL under Python 3.10 and trained on a single GPU. Training regimes differ by family: message-passing GNNs (GCN, GAT, GraphSAGE and their relational variants) use neighbor sampling (fan-out 10, 2 hops, batch size 10,240) and are evaluated with full-neighbor aggregation; HGT and SGFormer are trained over the full graph; SeHGNN and LMSPS precompute k-hop features and train an MLP head. Comparisons across model families hold the data and splits fixed, while sampling schemes follow each family's standard implementation. For each dataset–model pair, hyperparameters are held fixed across representations.

For CroVal, we train on questions dated up to 2015, validate on 2016–2018, and test from 2019 onward, yielding 9,699/9,889/14,565 nodes, respectively. This temporal split differs from the original benchmark's cutoffs of 2011/2012/2013 and is used consistently across all our CroVal experiments.

Experiments used NVIDIA RTX 3090 (24 GB), H100 PCIe (80 GB), and H100 (94 GB) GPUs, with one GPU per run. Configurations that ran out of memory on the RTX 3090 were retried on H100 GPUs. OOM denotes configurations that remained infeasible due to memory limits after these attempts. Such entries are excluded from aggregation rather than imputed.

Metrics. Both metrics are reported on a percentage scale for every configuration. Let TP_c, FP_c, FN_c be the counts for class c among C classes. Micro-F1 pools the counts before computing F1:

$$\mathrm{MicroF1} = \frac{2\sum_c \mathrm{TP}_c}{2\sum_c \mathrm{TP}_c + \sum_c \mathrm{FP}_c + \sum_c \mathrm{FN}_c}, \tag{4}$$

while Macro-F1 averages the per-class scores:

$$\mathrm{MacroF1} = \frac{1}{C}\sum_c \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{5}$$

Micro-F1 weights every prediction equally and is dominated by frequent classes; on the single-label datasets it coincides with accuracy. Macro-F1 weights every class equally and is therefore sensitive to minority classes. On the multi-label Book dataset, Micro-F1 aggregates over all node–label pairs and Macro-F1 averages label-wise binary F1 scores.

C Full Experimental Results

Table conventions. Standard deviations are given as subscripts. OOM marks configurations that exceeded available memory; these are excluded from aggregation rather than imputed. Random features are a topology-only control and are excluded from the column rankings. Within each column, bold marks the best and underline the second-best value, computed separately for Micro-F1 and Macro-F1; shading marks the top two by Micro-F1.

C.1 Full Representation Sweep

We evaluate MLP, SAGE, and RSAGE over all 15 representation tiers on all six datasets, holding the graph, labels, and data splits fixed. Tables 5–10 give the complete numbers used to compute text-only performance, $\Delta_{\mathrm{homo}} = \mathrm{SAGE} - \mathrm{MLP}$, and $\Delta_{\mathrm{rel}} = \mathrm{RSAGE} - \mathrm{SAGE}$. Random features are listed separately above the rule: they act as a topology-only control and are excluded from all regressions over text-derived representations, since they carry no text signal and would otherwise dominate the fit as a high-leverage outlier.

Table 5: Representation sweep on TMDB. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1.

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$25.22_{\pm 1.52}$ / $23.89_{\pm 1.10}$	$44.31_{\pm 2.04}$ / $44.93_{\pm 2.41}$	$46.06_{\pm 1.82}$ / $45.83_{\pm 1.54}$
Static	TF–IDF	$66.30_{\pm 0.17}$ / $66.01_{\pm 0.25}$	$76.28_{\pm 0.48}$ / $76.71_{\pm 0.57}$	$77.55_{\pm 0.45}$ / $78.37_{\pm 0.27}$
	GloVe	$70.88_{\pm 0.88}$ / $70.43_{\pm 0.89}$	$77.97_{\pm 0.88}$ / $78.70_{\pm 0.93}$	$79.40_{\pm 1.47}$ / $80.22_{\pm 1.29}$
PLM	DistilBERT	$74.09_{\pm 1.60}$ / $73.77_{\pm 1.47}$	$80.53_{\pm 0.74}$ / $81.15_{\pm 0.66}$	$83.41_{\pm 0.55}$ / $83.81_{\pm 0.47}$
	SimCSE	$75.31_{\pm 0.34}$ / $75.14_{\pm 0.28}$	$80.35_{\pm 0.24}$ / $81.22_{\pm 0.20}$	$83.45_{\pm 0.38}$ / $83.99_{\pm 0.41}$
	MPNet	$77.79_{\pm 0.44}$ / $77.31_{\pm 0.41}$	$82.55_{\pm 0.30}$ / $83.14_{\pm 0.26}$	$85.82_{\pm 0.37}$ / $86.29_{\pm 0.35}$
	SPECTER2	$61.77_{\pm 1.01}$ / $61.61_{\pm 0.99}$	$71.08_{\pm 1.34}$ / $71.96_{\pm 1.32}$	$74.20_{\pm 1.08}$ / $74.99_{\pm 0.91}$
	Orig. (MiniLM)	$72.19_{\pm 0.72}$ / $71.67_{\pm 0.78}$	$79.74_{\pm 0.59}$ / $80.40_{\pm 0.56}$	$82.46_{\pm 0.34}$ / $82.96_{\pm 0.32}$
	Sentence-T5	$80.53_{\pm 0.35}$ / $80.27_{\pm 0.41}$	$85.46_{\pm 0.47}$ / $85.67_{\pm 0.50}$	$85.77_{\pm 1.91}$ / $86.10_{\pm 1.96}$
Retrieval	BGE-large	$79.89_{\pm 0.16}$ / $79.41_{\pm 0.19}$	$83.23_{\pm 0.27}$ / $83.67_{\pm 0.28}$	$85.89_{\pm 0.78}$ / $86.15_{\pm 0.80}$
	BGE-M3	$77.21_{\pm 0.92}$ / $76.83_{\pm 0.78}$	$82.12_{\pm 0.26}$ / $82.86_{\pm 0.29}$	$84.49_{\pm 0.17}$ / $85.04_{\pm 0.19}$
LLM-based	Qwen3-0.6B	$77.12_{\pm 0.41}$ / $76.72_{\pm 0.46}$	$80.66_{\pm 0.50}$ / $81.41_{\pm 0.46}$	$84.58_{\pm 0.49}$ / $85.04_{\pm 0.49}$
	Qwen3-4B	$81.97_{\pm 0.41}$ / $81.79_{\pm 0.40}$	$85.46_{\pm 0.97}$ / $85.88_{\pm 0.88}$	$87.63_{\pm 0.54}$ / $87.95_{\pm 0.51}$
	Qwen3-8B	$\underline{83.89}_{\pm 0.50}$ / $\underline{83.77}_{\pm 0.47}$	$\underline{86.46}_{\pm 0.37}$ / $\underline{86.76}_{\pm 0.35}$	$\underline{88.54}_{\pm 0.44}$ / $\underline{88.88}_{\pm 0.41}$
	NV-Embed-v2	$\mathbf{87.63}_{\pm 0.36}$ / $\mathbf{87.55}_{\pm 0.41}$	$\mathbf{87.92}_{\pm 0.11}$ / $\mathbf{88.22}_{\pm 0.11}$	$\mathbf{91.00}_{\pm 0.36}$ / $\mathbf{91.20}_{\pm 0.32}$

Table 6: Representation sweep on CroVal. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1. The red cell is the DistilBERT–RSAGE configuration that collapsed to the majority class in all five seeds; it is excluded from the column rankings and from aggregate relation-gain statistics (see Figure 5).

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$46.65_{\pm 0.00}$ / $10.60_{\pm 0.00}$	$46.65_{\pm 0.00}$ / $10.60_{\pm 0.00}$	$46.64_{\pm 0.01}$ / $10.60_{\pm 0.00}$
Static	TF–IDF	$85.45_{\pm 0.14}$ / $81.86_{\pm 0.18}$	$85.80_{\pm 0.06}$ / $82.40_{\pm 0.12}$	$87.31_{\pm 0.09}$ / $84.10_{\pm 0.10}$
	GloVe	$80.76_{\pm 0.11}$ / $75.65_{\pm 0.19}$	$84.56_{\pm 0.16}$ / $80.35_{\pm 0.27}$	$85.41_{\pm 0.26}$ / $81.46_{\pm 0.48}$
PLM	DistilBERT	$77.81_{\pm 0.25}$ / $71.38_{\pm 0.44}$	$82.68_{\pm 0.21}$ / $77.81_{\pm 0.40}$	$47.22_{\pm 0.73}$ / $12.69_{\pm 2.58}$
	SimCSE	$82.11_{\pm 0.12}$ / $77.30_{\pm 0.20}$	$85.81_{\pm 0.08}$ / $82.02_{\pm 0.13}$	$85.82_{\pm 0.36}$ / $81.91_{\pm 0.49}$
	MPNet	$89.72_{\pm 0.07}$ / $87.24_{\pm 0.09}$	$88.26_{\pm 0.09}$ / $85.49_{\pm 0.10}$	$90.29_{\pm 0.05}$ / $87.93_{\pm 0.06}$
	SPECTER2	$88.79_{\pm 0.13}$ / $86.43_{\pm 0.17}$	$87.14_{\pm 0.11}$ / $84.09_{\pm 0.12}$	$89.41_{\pm 0.09}$ / $87.19_{\pm 0.14}$
	Orig. (MiniLM)	$88.62_{\pm 0.12}$ / $86.06_{\pm 0.13}$	$87.52_{\pm 0.08}$ / $84.69_{\pm 0.13}$	$89.38_{\pm 0.14}$ / $86.93_{\pm 0.15}$
	Sentence-T5	$88.81_{\pm 0.09}$ / $86.04_{\pm 0.15}$	$87.76_{\pm 0.19}$ / $84.86_{\pm 0.20}$	$89.72_{\pm 0.19}$ / $87.09_{\pm 0.28}$
Retrieval	BGE-large	$89.03_{\pm 0.04}$ / $86.65_{\pm 0.05}$	$88.17_{\pm 0.10}$ / $85.44_{\pm 0.15}$	$89.76_{\pm 0.13}$ / $87.40_{\pm 0.18}$
	BGE-M3	$86.02_{\pm 0.11}$ / $82.74_{\pm 0.11}$	$87.25_{\pm 0.07}$ / $84.30_{\pm 0.07}$	$87.68_{\pm 0.12}$ / $84.59_{\pm 0.09}$
LLM-based	Qwen3-0.6B	$90.19_{\pm 0.04}$ / $87.79_{\pm 0.08}$	$88.60_{\pm 0.09}$ / $85.85_{\pm 0.12}$	$90.65_{\pm 0.04}$ / $88.28_{\pm 0.07}$
	Qwen3-4B	$91.06_{\pm 0.05}$ / $89.06_{\pm 0.08}$	$89.04_{\pm 0.11}$ / $86.49_{\pm 0.14}$	$91.26_{\pm 0.12}$ / $89.14_{\pm 0.13}$
	Qwen3-8B	$\underline{91.24}_{\pm 0.07}$ / $\underline{89.27}_{\pm 0.09}$	$\underline{89.17}_{\pm 0.05}$ / $\underline{86.59}_{\pm 0.10}$	$\underline{91.32}_{\pm 0.10}$ / $\underline{89.35}_{\pm 0.12}$
	NV-Embed-v2	$\mathbf{91.84}_{\pm 0.07}$ / $\mathbf{90.03}_{\pm 0.08}$	$\mathbf{89.32}_{\pm 0.09}$ / $\mathbf{86.78}_{\pm 0.11}$	$\mathbf{91.84}_{\pm 0.08}$ / $\mathbf{90.05}_{\pm 0.12}$

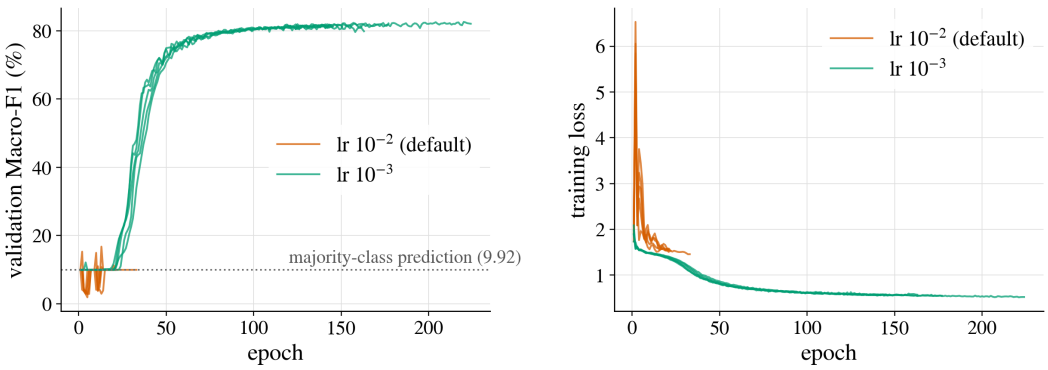


Figure 5: Training dynamics of RSAGE on CroVal with DistilBERT features over five seeds. At the shared learning rate of 10^{-2} (orange), all runs collapse to majority-class prediction. Changing only the learning rate to 10^{-3} (green) yields stable convergence and 83.02 ± 0.27 Micro-F1 / 77.93 ± 0.46 Macro-F1, confirming optimization instability rather than an information deficit. We retain the default-learning-rate result in Table 6 to preserve the fixed-hyperparameter comparison, but exclude this configuration from aggregate and correlation analyses involving Δ_{rel} or Δ_{best}. The retuned result is used only for this diagnostic and is not substituted into the main comparison.

Table 7: Representation sweep on arXiv. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1.

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$35.49_{\pm 0.00}$ / $1.34_{\pm 0.00}$	$48.23_{\pm 1.13}$ / $6.45_{\pm 1.20}$	$63.36_{\pm 1.25}$ / $21.37_{\pm 1.70}$
Static	TF–IDF	$73.22_{\pm 0.08}$ / $34.20_{\pm 0.43}$	$83.13_{\pm 0.32}$ / $47.93_{\pm 2.31}$	$82.82_{\pm 0.19}$ / $49.08_{\pm 0.81}$
	GloVe	$72.27_{\pm 0.31}$ / $34.95_{\pm 0.71}$	$81.59_{\pm 0.35}$ / $43.98_{\pm 1.73}$	$82.61_{\pm 0.27}$ / $48.05_{\pm 1.61}$
PLM	DistilBERT	$73.74_{\pm 0.37}$ / $36.20_{\pm 1.04}$	$81.00_{\pm 0.53}$ / $41.94_{\pm 1.36}$	$81.47_{\pm 0.65}$ / $46.35_{\pm 2.65}$
	SimCSE	$73.74_{\pm 0.18}$ / $36.05_{\pm 0.52}$	$82.14_{\pm 0.32}$ / $44.96_{\pm 1.60}$	$82.44_{\pm 0.42}$ / $45.91_{\pm 1.67}$
	MPNet	$80.50_{\pm 0.12}$ / $47.57_{\pm 2.64}$	$84.34_{\pm 0.10}$ / $\mathbf{53.15}_{\pm 1.95}$	$\mathbf{84.84}_{\pm 0.27}$ / $\mathbf{54.26}_{\pm 0.81}$
	SPECTER2	$80.49_{\pm 0.15}$ / $48.47_{\pm 1.50}$	$84.02_{\pm 0.33}$ / $49.62_{\pm 2.07}$	$84.57_{\pm 0.25}$ / $\underline{54.18}_{\pm 0.85}$
	Orig. (MiniLM)	$78.87_{\pm 0.12}$ / $43.33_{\pm 1.27}$	$83.97_{\pm 0.13}$ / $\underline{52.98}_{\pm 0.53}$	$\underline{84.65}_{\pm 0.13}$ / $53.98_{\pm 0.74}$
	Sentence-T5	$78.07_{\pm 0.25}$ / $41.99_{\pm 1.37}$	$83.16_{\pm 0.39}$ / $48.10_{\pm 0.78}$	$83.46_{\pm 0.38}$ / $50.82_{\pm 1.71}$
Retrieval	BGE-large	$80.10_{\pm 0.33}$ / $44.65_{\pm 1.99}$	$83.89_{\pm 0.19}$ / $50.02_{\pm 1.08}$	$84.47_{\pm 0.29}$ / $53.82_{\pm 1.19}$
	BGE-M3	$76.56_{\pm 0.08}$ / $41.02_{\pm 1.26}$	$83.04_{\pm 0.25}$ / $47.79_{\pm 2.01}$	$83.15_{\pm 0.34}$ / $51.94_{\pm 2.02}$
LLM-based	Qwen3-0.6B	$80.09_{\pm 0.20}$ / $47.09_{\pm 0.59}$	$84.01_{\pm 0.18}$ / $49.99_{\pm 1.10}$	$84.33_{\pm 0.25}$ / $52.71_{\pm 1.22}$
	Qwen3-4B	$81.72_{\pm 0.24}$ / $48.43_{\pm 2.60}$	$\mathbf{84.56}_{\pm 0.16}$ / $52.53_{\pm 2.14}$	$84.45_{\pm 0.09}$ / $53.81_{\pm 1.59}$
	Qwen3-8B	$\mathbf{82.19}_{\pm 0.18}$ / $\mathbf{49.22}_{\pm 1.41}$	$\underline{84.45}_{\pm 0.10}$ / $52.16_{\pm 1.51}$	$84.30_{\pm 0.10}$ / $53.99_{\pm 0.92}$
	NV-Embed-v2	$\underline{81.82}_{\pm 0.08}$ / $\underline{48.99}_{\pm 1.08}$	$84.41_{\pm 0.23}$ / $52.70_{\pm 2.62}$	$84.31_{\pm 0.20}$ / $53.57_{\pm 0.95}$

Table 8: Representation sweep on Book. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1.

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$21.26_{\pm 1.97}$ / $6.77_{\pm 1.40}$	$50.80_{\pm 0.68}$ / $31.09_{\pm 0.35}$	$55.16_{\pm 0.39}$ / $45.82_{\pm 0.54}$
Static	TF–IDF	$73.31_{\pm 0.19}$ / $64.34_{\pm 0.19}$	$81.82_{\pm 0.03}$ / $75.28_{\pm 0.04}$	$81.65_{\pm 0.07}$ / $75.78_{\pm 0.05}$
	GloVe	$75.06_{\pm 0.20}$ / $65.43_{\pm 0.61}$	$81.21_{\pm 0.06}$ / $74.20_{\pm 0.10}$	$82.29_{\pm 0.05}$ / $76.48_{\pm 0.10}$
PLM	DistilBERT	$79.15_{\pm 0.08}$ / $70.94_{\pm 0.40}$	$81.98_{\pm 0.05}$ / $74.93_{\pm 0.16}$	$83.46_{\pm 0.07}$ / $77.59_{\pm 0.14}$
	SimCSE	$78.69_{\pm 0.06}$ / $70.35_{\pm 0.24}$	$82.06_{\pm 0.09}$ / $75.28_{\pm 0.24}$	$83.28_{\pm 0.04}$ / $77.40_{\pm 0.13}$
	MPNet	$78.55_{\pm 0.10}$ / $70.66_{\pm 0.09}$	$82.27_{\pm 0.06}$ / $75.94_{\pm 0.11}$	$83.18_{\pm 0.07}$ / $77.63_{\pm 0.09}$
	SPECTER2	$71.66_{\pm 0.95}$ / $61.58_{\pm 2.57}$	$80.57_{\pm 0.09}$ / $73.45_{\pm 0.26}$	$81.20_{\pm 0.08}$ / $74.97_{\pm 0.18}$
	Orig. (MiniLM)	$75.07_{\pm 0.08}$ / $66.45_{\pm 0.26}$	$81.47_{\pm 0.04}$ / $74.97_{\pm 0.21}$	$82.14_{\pm 0.07}$ / $76.44_{\pm 0.12}$
	Sentence-T5	$81.01_{\pm 0.09}$ / $73.84_{\pm 0.30}$	$82.82_{\pm 0.07}$ / $76.38_{\pm 0.16}$	$84.54_{\pm 0.03}$ / $79.11_{\pm 0.08}$
Retrieval	BGE-large	$79.92_{\pm 0.05}$ / $72.83_{\pm 0.12}$	$82.68_{\pm 0.05}$ / $76.47_{\pm 0.18}$	$83.81_{\pm 0.07}$ / $78.45_{\pm 0.10}$
	BGE-M3	$78.06_{\pm 0.07}$ / $70.28_{\pm 0.21}$	$82.36_{\pm 0.05}$ / $75.93_{\pm 0.25}$	$83.03_{\pm 0.11}$ / $77.55_{\pm 0.10}$
LLM-based	Qwen3-0.6B	$79.26_{\pm 0.18}$ / $71.80_{\pm 0.15}$	$82.72_{\pm 0.04}$ / $76.37_{\pm 0.14}$	$83.65_{\pm 0.10}$ / $78.16_{\pm 0.09}$
	Qwen3-4B	$81.86_{\pm 0.12}$ / $75.42_{\pm 0.17}$	$83.59_{\pm 0.05}$ / $77.89_{\pm 0.09}$	$84.53_{\pm 0.15}$ / $79.31_{\pm 0.14}$
	Qwen3-8B	$\underline{82.59}_{\pm 0.11}$ / $\underline{76.51}_{\pm 0.08}$	$\underline{83.82}_{\pm 0.10}$ / $\underline{77.98}_{\pm 0.18}$	$\underline{84.67}_{\pm 0.08}$ / $\underline{79.51}_{\pm 0.10}$
	NV-Embed-v2	$\mathbf{84.10}_{\pm 0.15}$ / $\mathbf{78.41}_{\pm 0.14}$	$\mathbf{83.95}_{\pm 0.06}$ / $\mathbf{78.48}_{\pm 0.14}$	$\mathbf{85.38}_{\pm 0.14}$ / $\mathbf{80.37}_{\pm 0.17}$

Table 9: Representation sweep on DBLP. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1.

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$24.82_{\pm 0.00}$ / $4.42_{\pm 0.01}$	$27.22_{\pm 1.16}$ / $7.21_{\pm 1.09}$	$54.46_{\pm 0.46}$ / $48.12_{\pm 0.62}$
Static	TF–IDF	$66.99_{\pm 0.11}$ / $61.65_{\pm 0.18}$	$75.84_{\pm 0.16}$ / $72.59_{\pm 0.20}$	$75.66_{\pm 0.27}$ / $72.28_{\pm 0.49}$
	GloVe	$65.35_{\pm 0.02}$ / $59.69_{\pm 0.10}$	$75.02_{\pm 0.14}$ / $71.69_{\pm 0.18}$	$77.17_{\pm 0.21}$ / $74.09_{\pm 0.29}$
PLM	DistilBERT	$67.27_{\pm 0.09}$ / $62.40_{\pm 0.08}$	$75.05_{\pm 0.20}$ / $71.78_{\pm 0.25}$	$76.84_{\pm 0.36}$ / $73.85_{\pm 0.47}$
	SimCSE	$66.62_{\pm 0.07}$ / $61.59_{\pm 0.08}$	$75.28_{\pm 0.14}$ / $71.99_{\pm 0.21}$	$76.96_{\pm 0.26}$ / $73.99_{\pm 0.30}$
	MPNet	$71.05_{\pm 0.04}$ / $66.99_{\pm 0.10}$	$75.96_{\pm 0.23}$ / $72.91_{\pm 0.23}$	$\underline{77.47}_{\pm 0.13}$ / $\underline{74.55}_{\pm 0.22}$
	SPECTER2	$71.60_{\pm 0.09}$ / $67.71_{\pm 0.09}$	$75.94_{\pm 0.16}$ / $72.77_{\pm 0.22}$	$\mathbf{78.02}_{\pm 0.08}$ / $\mathbf{75.15}_{\pm 0.13}$
	Orig. (MiniLM)	$69.86_{\pm 0.07}$ / $65.41_{\pm 0.08}$	$75.82_{\pm 0.15}$ / $72.68_{\pm 0.16}$	$77.44_{\pm 0.23}$ / $74.52_{\pm 0.28}$
	Sentence-T5	$69.50_{\pm 0.11}$ / $64.95_{\pm 0.15}$	$75.38_{\pm 0.14}$ / $72.13_{\pm 0.17}$	$76.89_{\pm 0.29}$ / $73.86_{\pm 0.28}$
Retrieval	BGE-large	$71.46_{\pm 0.13}$ / $67.46_{\pm 0.12}$	$75.98_{\pm 0.12}$ / $72.84_{\pm 0.18}$	$77.34_{\pm 0.19}$ / $74.40_{\pm 0.28}$
	BGE-M3	$68.11_{\pm 0.06}$ / $63.29_{\pm 0.10}$	$75.53_{\pm 0.16}$ / $72.27_{\pm 0.21}$	$76.44_{\pm 0.23}$ / $73.31_{\pm 0.20}$
LLM-based	Qwen3-0.6B	$70.99_{\pm 0.03}$ / $67.01_{\pm 0.06}$	$76.05_{\pm 0.15}$ / $72.95_{\pm 0.20}$	$77.01_{\pm 0.11}$ / $74.01_{\pm 0.17}$
	Qwen3-4B	$72.42_{\pm 0.10}$ / $68.84_{\pm 0.15}$	$76.07_{\pm 0.22}$ / $\underline{73.03}_{\pm 0.30}$	$76.66_{\pm 0.21}$ / $73.45_{\pm 0.21}$
	Qwen3-8B	$\underline{72.70}_{\pm 0.08}$ / $\underline{69.12}_{\pm 0.13}$	$\underline{76.15}_{\pm 0.25}$ / $73.01_{\pm 0.30}$	$76.23_{\pm 0.20}$ / $73.05_{\pm 0.34}$
	NV-Embed-v2	$\mathbf{72.83}_{\pm 0.17}$ / $\mathbf{69.33}_{\pm 0.23}$	$\mathbf{76.26}_{\pm 0.09}$ / $\mathbf{73.21}_{\pm 0.19}$	$76.41_{\pm 0.18}$ / $73.28_{\pm 0.33}$

Table 10: Representation sweep on Patent. **Bold**/underline: best/second best per column, computed separately per metric; shading follows Micro-F1. The 4096-dimensional Qwen3-8B and NV-Embed-v2 features exceed memory for both message-passing models.

Type	Representation	MLP	SAGE	RSAGE
Control	Random	$13.26_{\pm 0.00}$ / $0.20_{\pm 0.00}$	$15.30_{\pm 0.03}$ / $0.44_{\pm 0.01}$	$15.55_{\pm 0.04}$ / $0.55_{\pm 0.06}$
Static	TF–IDF	$61.19_{\pm 0.07}$ / $33.45_{\pm 0.22}$	$73.06_{\pm 0.09}$ / $51.09_{\pm 0.23}$	$74.38_{\pm 0.17}$ / $52.92_{\pm 0.56}$
	GloVe	$63.99_{\pm 0.15}$ / $43.86_{\pm 0.18}$	$73.84_{\pm 0.16}$ / $56.11_{\pm 0.36}$	$75.59_{\pm 0.13}$ / $57.85_{\pm 0.20}$
PLM	DistilBERT	$66.38_{\pm 0.27}$ / $46.58_{\pm 0.57}$	$74.40_{\pm 0.03}$ / $56.68_{\pm 0.26}$	$75.80_{\pm 0.12}$ / $57.30_{\pm 0.29}$
	SimCSE	$67.21_{\pm 0.19}$ / $48.66_{\pm 0.26}$	$74.75_{\pm 0.13}$ / $57.91_{\pm 0.34}$	$76.15_{\pm 0.16}$ / $58.27_{\pm 0.48}$
	MPNet	$71.76_{\pm 0.15}$ / $55.79_{\pm 0.32}$	$76.38_{\pm 0.17}$ / $61.09_{\pm 0.57}$	$\underline{77.63}_{\pm 0.14}$ / $\underline{62.30}_{\pm 0.29}$
	SPECTER2	$69.56_{\pm 0.24}$ / $51.34_{\pm 0.52}$	$75.43_{\pm 0.26}$ / $58.66_{\pm 0.63}$	$76.79_{\pm 0.02}$ / $59.53_{\pm 0.48}$
	Orig. (MiniLM)	$69.64_{\pm 0.03}$ / $52.69_{\pm 0.16}$	$75.62_{\pm 0.10}$ / $59.80_{\pm 0.68}$	$76.98_{\pm 0.04}$ / $60.75_{\pm 0.25}$
	Sentence-T5	$71.96_{\pm 0.15}$ / $56.59_{\pm 0.19}$	$76.36_{\pm 0.17}$ / $61.20_{\pm 0.42}$	$77.55_{\pm 0.12}$ / $61.63_{\pm 0.33}$
Retrieval	BGE-large	$72.01_{\pm 0.17}$ / $56.62_{\pm 0.25}$	$76.44_{\pm 0.21}$ / $61.28_{\pm 0.41}$	$77.48_{\pm 0.18}$ / $61.97_{\pm 0.26}$
	BGE-M3	$69.97_{\pm 0.09}$ / $53.40_{\pm 0.31}$	$75.58_{\pm 0.14}$ / $59.58_{\pm 0.92}$	$76.90_{\pm 0.06}$ / $60.85_{\pm 0.11}$
LLM-based	Qwen3-0.6B	$72.56_{\pm 0.07}$ / $57.19_{\pm 0.23}$	$\underline{76.63}_{\pm 0.13}$ / $\underline{61.63}_{\pm 0.47}$	$\mathbf{77.79}_{\pm 0.10}$ / $\mathbf{62.68}_{\pm 0.43}$
	Qwen3-4B	$74.53_{\pm 0.06}$ / $60.63_{\pm 0.14}$	$\mathbf{77.14}_{\pm 0.14}$ / $\mathbf{63.03}_{\pm 0.29}$	$76.38_{\pm 0.17}$ / $59.52_{\pm 0.50}$
	Qwen3-8B	$\mathbf{74.90}_{\pm 0.12}$ / $\mathbf{61.32}_{\pm 0.15}$	OOM	OOM
	NV-Embed-v2	$\underline{74.87}_{\pm 0.09}$ / $\underline{61.24}_{\pm 0.20}$	OOM	OOM

C.2 Full Architecture Sweep

We compare 11 architectures across five representative node representations. Following the taxonomy in Section 5, the models cover text-only prediction, homogeneous message passing, relation-aware message passing, meta-path aggregation, and Transformer-based graph modeling. This sweep examines whether the representation-conditioned patterns observed in the controlled MLP–SAGE–RSAGE comparison extend across broader architectural choices. Each table reports Micro-F1 and Macro-F1 for all five representations and is presented in landscape orientation for readability. Ranks and averages are computed only over configurations completed by every model included in a given comparison. This ensures that differences in evaluation coverage do not affect the comparison.

All models take the same frozen text representation of target nodes as input and differ in which structural information they aggregate and how. Message-passing models, HGT, and SGFormer initialize auxiliary nodes with the mean of their target neighbors' features, whereas SeHGNN and LMSPS propagate target features along meta-paths. The homogeneous message-passing models (GCN, GAT, SAGE) and SGFormer operate on the type-agnostic union graph with self-loops. The relation-aware message-passing models (RGCN, RGAT, RSAGE) follow the implementation of Liu et al. (2025), averaging relation-specific outputs with equal weight and adding a linear skip connection on target nodes. Models are described below in table row order.

MLP. MLP applies a feed-forward transformation to each node's feature vector independently, with no exchange of information between nodes.

GCN. GCN aggregates degree-normalized neighbor representations through a first-order spectral filter, providing a simple form of untyped message passing (Kipf & Welling, 2016).

GAT. GAT uses multi-head attention to learn feature-dependent weights over neighbors rather than assigning weights solely from node degrees (Veličković et al., 2018).

GraphSAGE. SAGE (GraphSAGE) learns node representations by aggregating information from sampled neighborhoods (Hamilton et al., 2017). In our implementation, the node's own representation is averaged together with its neighbors' representations before a shared linear transformation.

RGCN. RGCN extends GCN to multi-relational graphs (Schlichtkrull et al., 2018). We apply a separate degree-normalized graph convolution for each relation type.

RGAT. RGAT is the relational analogue of GAT, computing attention coefficients under relation-specific transformations so that distinct weights are learned for each relation type.

RSAGE. RSAGE is the relational analogue of GraphSAGE, applying the same update separately for each relation type with relation-specific weights. The skip connection retains the target node's own representation as a separate channel.

SeHGNN. SeHGNN precomputes features aggregated along individual meta-paths and fuses the resulting semantic representations without iterative message passing (Yang et al., 2023b).

LMSPS. LMSPS controls the growth of long meta-paths on large heterogeneous graphs by progressively sampling and pruning candidate paths, then training on the selected subset (Li et al., 2024a).

HGT. HGT adapts Transformer attention to heterogeneous graphs through node-type-specific projections and relation-specific attention parameters (Hu et al., 2020).

SGFormer. SGFormer combines global node attention with a lightweight GCN branch while maintaining linear complexity in the number of nodes (Wu et al., 2023). Because it is designed for homogeneous graphs, we apply it to the type-agnostic union graph.

Table 11: Full architecture-sweep results on TMDB across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$72.19_{\pm 0.72}$	$71.67_{\pm 0.78}$	$77.79_{\pm 0.44}$	$77.31_{\pm 0.41}$	$79.89_{\pm 0.16}$	$79.41_{\pm 0.19}$	$83.89_{\pm 0.50}$	$83.77_{\pm 0.47}$	$87.63_{\pm 0.36}$	$87.55_{\pm 0.41}$
Homogeneous										
GCN	$77.74_{\pm 0.28}$	$78.57_{\pm 0.32}$	$80.53_{\pm 0.38}$	$81.16_{\pm 0.31}$	$81.51_{\pm 1.10}$	$81.98_{\pm 1.01}$	$82.59_{\pm 1.59}$	$82.96_{\pm 1.69}$	$83.45_{\pm 1.16}$	$83.72_{\pm 1.21}$
GAT	$79.12_{\pm 0.48}$	$79.88_{\pm 0.48}$	$81.35_{\pm 0.31}$	$82.06_{\pm 0.29}$	$83.43_{\pm 0.49}$	$83.86_{\pm 0.46}$	$86.48_{\pm 0.31}$	$86.80_{\pm 0.33}$	$87.12_{\pm 0.38}$	$87.35_{\pm 0.43}$
SAGE	$79.74_{\pm 0.59}$	$80.40_{\pm 0.56}$	$82.55_{\pm 0.30}$	$83.14_{\pm 0.26}$	$83.23_{\pm 0.27}$	$83.67_{\pm 0.28}$	$86.46_{\pm 0.37}$	$86.76_{\pm 0.35}$	$87.92_{\pm 0.11}$	$88.22_{\pm 0.11}$
Relation-aware										
RGCN	$81.57_{\pm 0.27}$	$82.13_{\pm 0.24}$	$\underline{85.47}_{\pm 0.26}$	$\underline{85.85}_{\pm 0.27}$	$86.09_{\pm 1.12}$	$86.37_{\pm 1.22}$	$88.28_{\pm 0.17}$	$88.56_{\pm 0.16}$	$\underline{90.77}_{\pm 0.39}$	$\underline{90.96}_{\pm 0.35}$
RGAT	$82.03_{\pm 0.41}$	$82.49_{\pm 0.38}$	$85.20_{\pm 0.41}$	$85.64_{\pm 0.39}$	$\underline{86.28}_{\pm 1.46}$	$\underline{86.58}_{\pm 1.35}$	$\mathbf{88.76}_{\pm 0.44}$	$\mathbf{89.03}_{\pm 0.41}$	$90.66_{\pm 0.36}$	$90.81_{\pm 0.37}$
RSAGE	$\mathbf{82.46}_{\pm 0.34}$	$\mathbf{82.96}_{\pm 0.32}$	$\mathbf{85.82}_{\pm 0.37}$	$\mathbf{86.29}_{\pm 0.35}$	$85.89_{\pm 0.78}$	$86.15_{\pm 0.80}$	$\underline{88.54}_{\pm 0.44}$	$\underline{88.88}_{\pm 0.41}$	$\mathbf{91.00}_{\pm 0.36}$	$\mathbf{91.20}_{\pm 0.32}$
Meta-path										
SeHGNN	$81.86_{\pm 0.59}$	$82.51_{\pm 0.60}$	$85.26_{\pm 0.29}$	$85.81_{\pm 0.30}$	$86.26_{\pm 0.12}$	$86.54_{\pm 0.14}$	$88.07_{\pm 0.34}$	$88.40_{\pm 0.32}$	$90.36_{\pm 0.29}$	$90.56_{\pm 0.27}$
LMSPS	$\underline{82.15}_{\pm 0.27}$	$\underline{82.75}_{\pm 0.25}$	$85.16_{\pm 0.20}$	$85.69_{\pm 0.19}$	$\mathbf{86.62}_{\pm 0.36}$	$\mathbf{86.94}_{\pm 0.38}$	$88.28_{\pm 0.27}$	$88.58_{\pm 0.27}$	$89.69_{\pm 0.32}$	$89.87_{\pm 0.31}$
Transformer-based										
HGT	$79.62_{\pm 0.63}$	$80.07_{\pm 0.60}$	$83.19_{\pm 1.14}$	$83.59_{\pm 1.14}$	$85.40_{\pm 0.54}$	$85.63_{\pm 0.49}$	$87.01_{\pm 1.00}$	$87.31_{\pm 0.98}$	$89.54_{\pm 0.30}$	$89.68_{\pm 0.35}$
SGFormer	$79.40_{\pm 0.89}$	$79.92_{\pm 0.83}$	$83.21_{\pm 0.54}$	$83.67_{\pm 0.47}$	$82.94_{\pm 0.57}$	$83.20_{\pm 0.49}$	$86.28_{\pm 0.42}$	$86.53_{\pm 0.42}$	$88.36_{\pm 0.35}$	$88.55_{\pm 0.36}$

Table 12: Full architecture-sweep results on CroVal across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$88.62_{\pm 0.12}$	$86.06_{\pm 0.13}$	$89.72_{\pm 0.07}$	$87.24_{\pm 0.09}$	$89.03_{\pm 0.04}$	$86.65_{\pm 0.05}$	$91.24_{\pm 0.07}$	$89.27_{\pm 0.09}$	$91.84_{\pm 0.07}$	$90.03_{\pm 0.08}$
Homogeneous										
GCN	$85.33_{\pm 0.12}$	$81.79_{\pm 0.14}$	$85.95_{\pm 0.12}$	$82.53_{\pm 0.17}$	$85.82_{\pm 0.13}$	$82.29_{\pm 0.11}$	$86.71_{\pm 0.16}$	$83.45_{\pm 0.24}$	$86.79_{\pm 0.16}$	$83.51_{\pm 0.16}$
GAT	$86.87_{\pm 0.35}$	$83.77_{\pm 0.35}$	$87.92_{\pm 0.69}$	$85.01_{\pm 0.75}$	$87.82_{\pm 0.89}$	$84.76_{\pm 0.97}$	$89.37_{\pm 0.16}$	$86.72_{\pm 0.23}$	$89.21_{\pm 0.38}$	$86.59_{\pm 0.39}$
SAGE	$87.52_{\pm 0.08}$	$84.69_{\pm 0.13}$	$88.26_{\pm 0.09}$	$85.49_{\pm 0.10}$	$88.17_{\pm 0.10}$	$85.44_{\pm 0.15}$	$89.17_{\pm 0.05}$	$86.59_{\pm 0.10}$	$89.32_{\pm 0.09}$	$86.78_{\pm 0.11}$
Relation-aware										
RGCN	$89.31_{\pm 0.10}$	$86.81_{\pm 0.13}$	$\underline{90.35}_{\pm 0.06}$	$\underline{88.04}_{\pm 0.10}$	$\underline{89.94}_{\pm 0.08}$	$\mathbf{87.66}_{\pm 0.15}$	$\mathbf{91.47}_{\pm 0.14}$	$\mathbf{89.53}_{\pm 0.20}$	$\underline{91.87}_{\pm 0.10}$	$\underline{90.13}_{\pm 0.17}$
RGAT	$\mathbf{89.43}_{\pm 0.11}$	$\mathbf{87.01}_{\pm 0.12}$	$\mathbf{90.40}_{\pm 0.15}$	$\mathbf{88.08}_{\pm 0.21}$	$\mathbf{89.95}_{\pm 0.06}$	$\underline{87.64}_{\pm 0.10}$	$\underline{91.44}_{\pm 0.06}$	$\underline{89.52}_{\pm 0.08}$	$\mathbf{91.96}_{\pm 0.09}$	$\mathbf{90.22}_{\pm 0.16}$
RSAGE	$\underline{89.38}_{\pm 0.14}$	$\underline{86.93}_{\pm 0.15}$	$90.29_{\pm 0.05}$	$87.93_{\pm 0.06}$	$89.76_{\pm 0.13}$	$87.40_{\pm 0.18}$	$91.32_{\pm 0.10}$	$89.35_{\pm 0.12}$	$91.84_{\pm 0.08}$	$90.05_{\pm 0.12}$
Meta-path										
SeHGNN	$88.30_{\pm 0.11}$	$85.21_{\pm 0.15}$	$89.38_{\pm 0.10}$	$86.59_{\pm 0.14}$	$89.34_{\pm 0.10}$	$86.51_{\pm 0.14}$	$90.74_{\pm 0.08}$	$88.43_{\pm 0.15}$	$91.29_{\pm 0.15}$	$89.01_{\pm 0.23}$
LMSPS	$88.00_{\pm 0.16}$	$84.75_{\pm 0.23}$	$88.99_{\pm 0.04}$	$85.99_{\pm 0.08}$	$88.99_{\pm 0.05}$	$86.02_{\pm 0.07}$	$90.57_{\pm 0.15}$	$88.09_{\pm 0.13}$	$90.97_{\pm 0.11}$	$88.55_{\pm 0.16}$
Transformer-based										
HGT	$89.26_{\pm 0.07}$	$86.64_{\pm 0.12}$	$90.14_{\pm 0.10}$	$87.59_{\pm 0.14}$	$88.98_{\pm 0.08}$	$86.46_{\pm 0.09}$	$90.86_{\pm 0.08}$	$88.79_{\pm 0.12}$	$91.72_{\pm 0.14}$	$89.82_{\pm 0.18}$
SGFormer	$89.16_{\pm 0.16}$	$86.77_{\pm 0.20}$	$89.98_{\pm 0.08}$	$87.60_{\pm 0.09}$	$89.15_{\pm 0.24}$	$86.85_{\pm 0.28}$	$90.69_{\pm 0.08}$	$88.73_{\pm 0.12}$	$91.20_{\pm 0.14}$	$89.17_{\pm 0.25}$

Table 13: Full architecture-sweep results on arXiv across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$78.87_{\pm 0.12}$	$43.33_{\pm 1.27}$	$80.50_{\pm 0.12}$	$47.57_{\pm 2.64}$	$80.10_{\pm 0.33}$	$44.65_{\pm 1.99}$	$82.19_{\pm 0.18}$	$49.22_{\pm 1.41}$	$81.82_{\pm 0.08}$	$48.99_{\pm 1.08}$
Homogeneous										
GCN	$82.05_{\pm 0.20}$	$50.11_{\pm 1.36}$	$82.44_{\pm 0.07}$	$49.56_{\pm 1.22}$	$82.09_{\pm 0.15}$	$48.59_{\pm 1.84}$	$82.64_{\pm 0.24}$	$49.75_{\pm 2.43}$	$82.67_{\pm 0.19}$	$49.73_{\pm 1.99}$
GAT	$83.97_{\pm 0.25}$	$50.27_{\pm 1.74}$	$84.36_{\pm 0.25}$	$50.65_{\pm 1.62}$	$83.94_{\pm 0.25}$	$48.89_{\pm 1.25}$	$\mathbf{84.75}_{\pm 0.23}$	$53.19_{\pm 1.15}$	$\mathbf{84.63}_{\pm 0.21}$	$50.89_{\pm 2.02}$
SAGE	$83.97_{\pm 0.13}$	$52.98_{\pm 0.53}$	$84.34_{\pm 0.10}$	$53.15_{\pm 1.95}$	$83.89_{\pm 0.19}$	$50.02_{\pm 1.08}$	$84.45_{\pm 0.10}$	$52.16_{\pm 1.51}$	$84.41_{\pm 0.23}$	$52.70_{\pm 2.62}$
Relation-aware										
RGCN	$84.53_{\pm 0.23}$	$52.59_{\pm 1.06}$	$84.73_{\pm 0.19}$	$53.60_{\pm 1.69}$	$\mathbf{84.55}_{\pm 0.23}$	$53.79_{\pm 1.38}$	$84.51_{\pm 0.13}$	$\underline{53.20}_{\pm 1.17}$	$84.27_{\pm 0.08}$	$\underline{53.30}_{\pm 0.93}$
RGAT	$\underline{84.60}_{\pm 0.26}$	$\underline{53.06}_{\pm 0.91}$	$\mathbf{84.89}_{\pm 0.21}$	$\underline{53.76}_{\pm 0.88}$	$84.40_{\pm 0.21}$	$\mathbf{54.18}_{\pm 0.56}$	$\underline{84.72}_{\pm 0.12}$	$53.04_{\pm 0.72}$	$\underline{84.54}_{\pm 0.22}$	$53.01_{\pm 0.77}$
RSAGE	$\mathbf{84.65}_{\pm 0.13}$	$\mathbf{53.98}_{\pm 0.74}$	$\underline{84.84}_{\pm 0.27}$	$\mathbf{54.26}_{\pm 0.81}$	$\underline{84.47}_{\pm 0.29}$	$\underline{53.82}_{\pm 1.19}$	$84.30_{\pm 0.10}$	$\mathbf{53.99}_{\pm 0.92}$	$84.31_{\pm 0.20}$	$\mathbf{53.57}_{\pm 0.95}$
Meta-path										
SeHGNN	$81.74_{\pm 0.17}$	$44.57_{\pm 0.48}$	$82.23_{\pm 0.18}$	$45.50_{\pm 0.46}$	$82.26_{\pm 0.37}$	$44.93_{\pm 0.39}$	$82.72_{\pm 0.09}$	$45.08_{\pm 0.62}$	$82.63_{\pm 0.15}$	$45.02_{\pm 0.41}$
LMSPS	$80.65_{\pm 0.18}$	$45.94_{\pm 0.21}$	$81.07_{\pm 0.30}$	$46.93_{\pm 0.46}$	$80.37_{\pm 0.39}$	$45.48_{\pm 0.87}$	$80.58_{\pm 0.18}$	$45.11_{\pm 0.56}$	$80.88_{\pm 0.30}$	$45.63_{\pm 0.77}$
Transformer-based										
HGT	$81.53_{\pm 0.14}$	$44.99_{\pm 0.60}$	$81.94_{\pm 0.22}$	$46.21_{\pm 1.91}$	$81.40_{\pm 0.08}$	$45.64_{\pm 1.22}$	$82.47_{\pm 0.13}$	$46.54_{\pm 2.10}$	$82.20_{\pm 0.12}$	$47.19_{\pm 1.12}$
SGFormer	$82.26_{\pm 0.38}$	$49.74_{\pm 1.01}$	$82.30_{\pm 0.39}$	$50.17_{\pm 0.29}$	$81.75_{\pm 0.70}$	$50.30_{\pm 1.62}$	$82.15_{\pm 0.45}$	$51.86_{\pm 0.95}$	$82.06_{\pm 0.36}$	$51.54_{\pm 1.11}$

Table 14: Full architecture-sweep results on Book across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column. HGT exceeds memory at every tier on this dataset.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$75.07_{\pm 0.08}$	$66.45_{\pm 0.26}$	$78.55_{\pm 0.10}$	$70.66_{\pm 0.09}$	$79.92_{\pm 0.05}$	$72.83_{\pm 0.12}$	$82.59_{\pm 0.11}$	$76.51_{\pm 0.08}$	$84.10_{\pm 0.15}$	$78.41_{\pm 0.14}$
Homogeneous										
GCN	$80.71_{\pm 0.12}$	$73.95_{\pm 0.15}$	$81.67_{\pm 0.06}$	$75.64_{\pm 0.11}$	$82.06_{\pm 0.09}$	$75.93_{\pm 0.18}$	$82.94_{\pm 0.08}$	$76.67_{\pm 0.28}$	$83.15_{\pm 0.05}$	$77.14_{\pm 0.17}$
GAT	$80.98_{\pm 0.06}$	$74.19_{\pm 0.18}$	$81.92_{\pm 0.03}$	$75.47_{\pm 0.06}$	$82.33_{\pm 0.08}$	$75.72_{\pm 0.10}$	$83.42_{\pm 0.07}$	$77.41_{\pm 0.25}$	$83.69_{\pm 0.05}$	$77.81_{\pm 0.04}$
SAGE	$81.47_{\pm 0.04}$	$74.97_{\pm 0.21}$	$82.27_{\pm 0.06}$	$75.94_{\pm 0.11}$	$82.68_{\pm 0.05}$	$76.47_{\pm 0.18}$	$83.82_{\pm 0.10}$	$77.98_{\pm 0.18}$	$83.95_{\pm 0.06}$	$78.48_{\pm 0.14}$
Relation-aware										
RGCN	$82.04_{\pm 0.07}$	$76.36_{\pm 0.06}$	$83.10_{\pm 0.14}$	$77.60_{\pm 0.16}$	$83.78_{\pm 0.04}$	$78.36_{\pm 0.01}$	$84.57_{\pm 0.08}$	$79.38_{\pm 0.08}$	$85.31_{\pm 0.11}$	$80.34_{\pm 0.08}$
RGAT	$\underline{82.32}_{\pm 0.03}$	$\underline{76.67}_{\pm 0.08}$	$\underline{83.43}_{\pm 0.03}$	$\underline{78.02}_{\pm 0.06}$	$\underline{84.10}_{\pm 0.05}$	$\underline{78.77}_{\pm 0.05}$	$\underline{85.12}_{\pm 0.06}$	$\underline{80.08}_{\pm 0.10}$	$\mathbf{85.64}_{\pm 0.04}$	$\mathbf{80.75}_{\pm 0.04}$
RSAGE	$82.14_{\pm 0.07}$	$76.44_{\pm 0.12}$	$83.18_{\pm 0.07}$	$77.63_{\pm 0.09}$	$83.81_{\pm 0.07}$	$78.45_{\pm 0.10}$	$84.67_{\pm 0.08}$	$79.51_{\pm 0.10}$	$\underline{85.38}_{\pm 0.14}$	$80.37_{\pm 0.17}$
Meta-path										
SeHGNN	$81.45_{\pm 0.14}$	$73.39_{\pm 0.31}$	$82.38_{\pm 0.13}$	$74.25_{\pm 0.22}$	$83.01_{\pm 0.12}$	$75.05_{\pm 0.25}$	$84.31_{\pm 0.07}$	$76.89_{\pm 0.14}$	$84.83_{\pm 0.05}$	$77.88_{\pm 0.14}$
LMSPS	$78.92_{\pm 0.25}$	$69.70_{\pm 0.30}$	$80.14_{\pm 0.19}$	$70.96_{\pm 0.26}$	$80.74_{\pm 0.24}$	$71.44_{\pm 0.34}$	OOM		OOM	
Transformer-based										
HGT	OOM		OOM		OOM		OOM		OOM	
SGFormer	$\mathbf{83.04}_{\pm 0.07}$	$\mathbf{77.36}_{\pm 0.07}$	$\mathbf{83.99}_{\pm 0.08}$	$\mathbf{78.42}_{\pm 0.07}$	$\mathbf{84.47}_{\pm 0.07}$	$\mathbf{79.16}_{\pm 0.07}$	$\mathbf{85.20}_{\pm 0.03}$	$\mathbf{80.22}_{\pm 0.03}$	$85.35_{\pm 0.09}$	$\underline{80.44}_{\pm 0.12}$

Table 15: Full architecture-sweep results on DBLP across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$69.86_{\pm 0.07}$	$65.41_{\pm 0.08}$	$71.05_{\pm 0.04}$	$66.99_{\pm 0.10}$	$71.46_{\pm 0.13}$	$67.46_{\pm 0.12}$	$72.70_{\pm 0.08}$	$69.12_{\pm 0.13}$	$72.83_{\pm 0.17}$	$69.33_{\pm 0.23}$
Homogeneous										
GCN	$73.71_{\pm 0.17}$	$70.11_{\pm 0.21}$	$74.00_{\pm 0.08}$	$70.36_{\pm 0.19}$	$74.30_{\pm 0.13}$	$70.83_{\pm 0.11}$	$74.59_{\pm 0.21}$	$71.10_{\pm 0.25}$	$74.53_{\pm 0.22}$	$71.05_{\pm 0.30}$
GAT	$75.49_{\pm 0.12}$	$72.32_{\pm 0.15}$	$75.83_{\pm 0.18}$	$72.80_{\pm 0.19}$	$75.81_{\pm 0.12}$	$72.73_{\pm 0.19}$	$76.03_{\pm 0.10}$	$72.96_{\pm 0.25}$	$76.16_{\pm 0.24}$	$73.16_{\pm 0.30}$
SAGE	$75.82_{\pm 0.15}$	$72.68_{\pm 0.16}$	$75.96_{\pm 0.23}$	$72.91_{\pm 0.23}$	$75.98_{\pm 0.12}$	$72.84_{\pm 0.18}$	$76.15_{\pm 0.25}$	$73.01_{\pm 0.30}$	$76.26_{\pm 0.09}$	$73.21_{\pm 0.19}$
Relation-aware										
RGCN	$77.26_{\pm 0.26}$	$74.32_{\pm 0.33}$	$76.95_{\pm 0.12}$	$73.93_{\pm 0.22}$	$77.09_{\pm 0.23}$	$73.97_{\pm 0.39}$	$\underline{76.25}_{\pm 0.10}$	$\underline{73.17}_{\pm 0.19}$	$76.31_{\pm 0.23}$	$73.12_{\pm 0.44}$
RGAT	$\mathbf{77.87}_{\pm 0.21}$	$\mathbf{74.95}_{\pm 0.20}$	$\mathbf{77.89}_{\pm 0.14}$	$\mathbf{74.96}_{\pm 0.25}$	$\mathbf{77.37}_{\pm 0.11}$	$\underline{74.25}_{\pm 0.19}$	$\mathbf{77.25}_{\pm 0.26}$	$\mathbf{74.26}_{\pm 0.33}$	$\mathbf{77.20}_{\pm 0.13}$	$\mathbf{74.08}_{\pm 0.20}$
RSAGE	$\underline{77.44}_{\pm 0.23}$	$\underline{74.52}_{\pm 0.28}$	$\underline{77.47}_{\pm 0.13}$	$\underline{74.55}_{\pm 0.22}$	$\underline{77.34}_{\pm 0.19}$	$\mathbf{74.40}_{\pm 0.28}$	$76.23_{\pm 0.20}$	$73.05_{\pm 0.34}$	$\underline{76.41}_{\pm 0.18}$	$\underline{73.28}_{\pm 0.33}$
Meta-path										
SeHGNN	$76.42_{\pm 0.08}$	$73.30_{\pm 0.07}$	$76.35_{\pm 0.08}$	$73.27_{\pm 0.16}$	$76.44_{\pm 0.13}$	$73.39_{\pm 0.19}$	OOM		OOM	
LMSPS	$76.18_{\pm 0.08}$	$73.19_{\pm 0.17}$	OOM		OOM		OOM		OOM	
Transformer-based										
HGT	OOM		OOM		OOM		OOM		OOM	
SGFormer	$75.45_{\pm 0.13}$	$72.52_{\pm 0.11}$	$75.34_{\pm 0.09}$	$72.52_{\pm 0.15}$	$75.56_{\pm 0.11}$	$72.66_{\pm 0.08}$	OOM		OOM	

Table 16: Full architecture-sweep results on Patent across five representative representation tiers. Each entry is the mean over five seeds with the standard deviation as a subscript. Row shading groups models by inductive-bias family; **bold** marks the best and underline the second-best value per column. Patent is the largest graph in the benchmark: every message-passing model exceeds memory at the two 4096-dimensional tiers, leaving only the text-only baseline available there.

	Orig. (MiniLM)		MPNet		BGE-large		Qwen3-8B		NV-Embed-v2	
Model	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Text-only										
MLP	$69.64_{\pm 0.03}$	$52.69_{\pm 0.16}$	$71.76_{\pm 0.15}$	$55.79_{\pm 0.32}$	$72.01_{\pm 0.17}$	$56.62_{\pm 0.25}$	**74.90**$_{\pm 0.12}$	**61.32**$_{\pm 0.15}$	**74.87**$_{\pm 0.09}$	**61.24**$_{\pm 0.20}$
Homogeneous										
GCN	$73.63_{\pm 0.14}$	$56.55_{\pm 0.37}$	$74.29_{\pm 0.33}$	$57.92_{\pm 0.41}$	$74.00_{\pm 0.30}$	$57.38_{\pm 0.86}$	OOM		OOM	
GAT	$74.68_{\pm 0.22}$	$56.20_{\pm 0.71}$	$75.43_{\pm 0.16}$	$57.67_{\pm 0.17}$	$75.66_{\pm 0.17}$	$58.09_{\pm 0.19}$	OOM		OOM	
SAGE	$75.62_{\pm 0.10}$	$59.80_{\pm 0.68}$	$76.38_{\pm 0.17}$	$61.09_{\pm 0.57}$	$76.44_{\pm 0.21}$	$61.28_{\pm 0.41}$	OOM		OOM	
Relation-aware										
RGCN	$76.90_{\pm 0.08}$	$60.52_{\pm 0.49}$	$77.52_{\pm 0.11}$	$61.44_{\pm 0.42}$	$\underline{77.49}_{\pm 0.13}$	$\underline{62.04}_{\pm 0.18}$	OOM		OOM	
RGAT	**77.20**$_{\pm 0.12}$	**61.29**$_{\pm 0.23}$	**77.88**$_{\pm 0.10}$	**62.50**$_{\pm 0.07}$	**77.73**$_{\pm 0.19}$	**62.19**$_{\pm 0.45}$	OOM		OOM	
RSAGE	$76.98_{\pm 0.04}$	$\underline{60.75}_{\pm 0.25}$	$\underline{77.63}_{\pm 0.14}$	$\underline{62.30}_{\pm 0.29}$	$77.48_{\pm 0.18}$	$61.97_{\pm 0.26}$	OOM		OOM	
Meta-path										
SeHGNN	$76.35_{\pm 0.11}$	$54.58_{\pm 0.17}$	$76.93_{\pm 0.09}$	$56.79_{\pm 0.14}$	$77.21_{\pm 0.17}$	$56.81_{\pm 0.21}$	OOM		OOM	
LMSPS	$\underline{77.13}_{\pm 0.11}$	$58.42_{\pm 0.37}$	OOM		OOM		OOM		OOM	
Transformer-based										
HGT	OOM		OOM		OOM		OOM		OOM	
SGFormer	OOM		OOM		OOM		OOM		OOM	

D ADDITIONAL ANALYSES OF TGA

This section collects the computational details and supplementary analyses of TGA. We first list the enumerated target-returning meta-paths and describe the pair-sampling and computation procedures. We then derive the label-only null introduced in Section 6.1 and report how the empirical TGA landscape departs from it. Next, we examine sampling stability and show that TGA is calibrated within, but not across, HTAGs. Finally, we show that the TGA matrix is approximately rank one, with representation-side and graph-side factors that explain most of its within-HTAG ordering.

D.1 ENUMERATED META-PATHS

We enumerate all target-returning meta-paths of up to four hops over the undirected type schema and deduplicate them by node-type sequence. Table 17 lists the complete resulting sets. No enumerated path is filtered: every listed path yields at least one non-self target-node pair.

Table 17: Enumerated target-returning meta-paths used for TGA. All schema-valid paths of up to four hops are retained.

Dataset	Target-returning meta-paths	# paths
TMDB	MAM, MDM, MAMAM, MAMDM, MDMAM, MDMDM	6
CroVal	QQ, QQQ, QTQ, QUQ, QQQQ, QQTQ, QQUQ, QTQQ, QUQQ, QQQQQ, QQQTQ, QQQUQ, QQTQQ, QQUQQ, QTQQQ, QTQTQ, QTQUQ, QUQQQ, QUQTQ, QUQUQ	20
arXiv	PP, PAP, PFP, PPP, PAPP, PFPP, PPAP, PPFP, PPPP, PAPAP, PAPFP, PAPPP, PFPAP, PFPFP, PFPPP, PPAPP, PPFPP, PPPAP, PPPFP, PPPPP	20
Book	BB, BAB, BBB, BUB, BABB, BBAB, BBBB, BBUB, BUBB, BABAB, BABBB, BABUB, BBABB, BBBAB, BBBBB, BBBUB, BBUBB, BUBAB, BUBBB, BUBUB	20
DBLP	PP, PAP, PFP, PPP, PAPP, PFPP, PPAP, PPFP, PPPP, PAPAP, PAPFP, PAPPP, PFPAP, PFPFP, PFPPP, PPAPP, PPFPP, PPPAP, PPPFP, PPPPP	20
Patent	PEP, PIP, PEPEP, PEPIP, PIPEP, PIPIP	6
Total		92

Abbreviations. M: movie; A: actor (TMDB) or author (arXiv, Book, and DBLP); D: director; Q: question; U: user (CroVal) or publisher (Book); T: tag; P: paper (arXiv and DBLP) or patent (Patent); F: field of study; B: book; E: examiner; I: inventor.

D.2 TGA COMPUTATION

Connected and random pair construction. For each target-returning meta-path m, we sample walks that follow the typed relation sequence specified by m. Each successful walk produces a pair (u, v) consisting of its target-type start and end nodes. Walks that cannot complete the prescribed relation sequence are discarded, as are self-returning walks with $u = v$. The retained pairs form the connected-pair set $\mathcal{P}_m$.

For every $\mathcal{P}_m$, we construct an equal-sized reference set $\mathcal{R}_m$ by independently sampling pairs of target nodes. This preserves the target node type while removing the connectivity constraint induced by the meta-path. We attempt ($K = 10{,}000$) walks per meta-path and retain only successful, non-self-returning pairs, so the retained count can be smaller than (K).

Given representation e, let

$$s_m^+(u, v) = \cos\left(\mathbf{x}_u^{(e)}, \mathbf{x}_v^{(e)}\right), \qquad (u, v) \in \mathcal{P}_m, \tag{6}$$

and define s_m^- analogously over $\mathcal{R}_m$. TGA compares the two similarity distributions through

$$\mathrm{TGA}(e, m) = \Pr(s_m^+ > s_m^-) + \tfrac{1}{2}\Pr(s_m^+ = s_m^-). \tag{7}$$

This is the normalized Mann–Whitney U statistic, or equivalently the AUC obtained when connected and independently paired node pairs are discriminated using representation similarity alone. A value of 0.5 indicates no ordering between the two distributions, while values above 0.5 indicate that meta-path-connected nodes tend to be more similar than independently paired target nodes.

Table 18: **TGA sampling statistics.** Number of valid target-returning meta-paths and average number of retained connected pairs per meta-path after failed and self-returning walks are removed. Sampling uses (K=10,000) walk attempts per meta-path.

Dataset	Valid paths	Avg. retained pairs
TMDB	6	8,957
CroVal	20	3,746
arXiv	20	8,884
Book	20	9,219
DBLP	20	9,617
Patent	6	8,516

Dataset-level TGA is obtained by macro-averaging over the valid target-returning meta-paths,

$$\mathrm{TGA}(e) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathrm{TGA}(e, m), \tag{8}$$

where $\mathcal{M}$ denotes the set of valid meta-paths for that HTAG. Each meta-path therefore contributes equally regardless of the number of retained walks.

Efficient computation. A direct implementation of equation 7 would compare every connected-pair similarity against every random-pair similarity and require quadratic time in the number of sampled pairs. Instead, we sort the random similarities and use binary search to count how many are smaller than or equal to each connected similarity. This computes the same Mann–Whitney statistic in $O(K \log K)$ time and avoids materializing a $K \times K$ comparison matrix.

Table 18 summarizes the realized sampling statistics. The smaller retained count on CroVal is primarily due to failed walks. Of all attempted CroVal walks, 58.2% cannot complete their prescribed relation sequence, while 4.4% complete it but return to the starting node.

D.3 Derivation of the Label-Only Null

Section 6.1 introduces a deliberately restrictive null in which the effect of connectivity on representation similarity is mediated entirely through the observed class labels. We provide the derivation here.

Let

$$h_m = \Pr(y_u = y_v \mid (u, v) \in \mathcal{P}_m) \tag{9}$$

denote label agreement among meta-path-connected pairs. For uniformly sampled distinct target-node pairs, the corresponding agreement probability is

$$h_0 = \frac{N \sum_c \pi_c^2 - 1}{N - 1}, \tag{10}$$

where N is the number of target nodes and π_c is the target-node class prior.

Let S^{same} and S^{diff} denote similarity random variables for same-class and different-class node pairs. We define the class separability of representation e as

$$A(e) = \Pr(S^{\mathrm{same}} > S^{\mathrm{diff}}) + \tfrac{1}{2} \Pr(S^{\mathrm{same}} = S^{\mathrm{diff}}). \tag{11}$$

The null relies on two assumptions. Assumption (A1) states that, conditional on whether the labels agree, connected and random pairs follow the same similarity distributions. Assumption (A2) states that the class composition within the same-label and different-label events does not otherwise differ between the connected and random pair populations. Under these assumptions, the positive and negative similarity distributions can be written as

$$S_m^+ \sim h_m S^{\mathrm{same}} + (1 - h_m) S^{\mathrm{diff}}, \tag{12}$$

$$S_m^- \sim h_0 S^{\mathrm{same}} + (1 - h_0) S^{\mathrm{diff}}. \tag{13}$$

Table 19: **TGA sampling stability on TMDB.** Mean and standard deviation over 20 independent resamplings.

Configuration	Mean TGA	Std. dev.
Qwen3-0.6B@1024	0.6102	0.0025
Qwen3-4B@1024	0.6221	0.0022
Qwen3-8B@1024	0.6298	0.0019
Qwen3-8B@2560	0.6321	0.0020
Qwen3-8B@4096	0.6345	0.0020

Expanding the probability that a draw from S_m^+ exceeds a draw from S_m^- gives four cases. Comparisons between two draws from the same conditional distribution contribute $\frac{1}{2}$ by symmetry. Same-class versus different-class comparisons contribute $A(e)$, while the reverse comparison contributes $1 - A(e)$. Therefore,

$$\begin{aligned}\mathrm{TGA}(e, m) = \tfrac{1}{2}h_m h_0 + A(e)h_m(1 - h_0) \\ + \big(1 - A(e)\big)(1 - h_m)h_0 + \tfrac{1}{2}(1 - h_m)(1 - h_0).\end{aligned} \tag{14}$$

Collecting terms yields

$$\mathrm{TGA}(e, m) = \tfrac{1}{2} + \Big(A(e) - \tfrac{1}{2}\Big)\Big(h_m - h_0\Big), \tag{15}$$

which recovers equation 3.

When $h_m \geq h_0$, the constraint $A(e) \leq 1$ further gives

$$\mathrm{TGA}(e, m) \leq \tfrac{1}{2} + \tfrac{1}{2}(h_m - h_0). \tag{16}$$

This is an upper bound only under the joint label-only assumptions. It is not intended as a generative model of real HTAGs.

D.4 Empirical Departures from the Label-Only Null

The exact label-only null is strongly inconsistent with the empirical TGA landscape. Across the 1,288 text-derived representation–meta-path cells considered in the analysis, 89% (1,151 cells) exceed the upper bound in equation 16. The observed alignment therefore exceeds what the combination of label homophily and representation-level class separability can produce under assumptions (A1) and (A2).

This violation should be interpreted carefully. Because the bound follows from the two assumptions jointly, exceeding it rejects their conjunction. It does not by itself identify whether the discrepancy arises from connectivity-dependent similarity within a class pair, from differences in class composition, or from both. The result should therefore be read as evidence that the coefficient-one label-only model systematically underestimates the alignment measured by TGA, rather than as evidence that TGA is unrelated to label structure.

D.5 Sampling Stability and Within-HTAG Calibration

Sampling Stability TGA is computed from sampled node pairs rather than from all possible pairs, so we also examine whether the measured differences are large relative to sampling variation. We repeat the TMDB computation 20 times with independent pair samples for representative dimension-controlled Qwen3 configurations. Table 19 reports the resulting means and standard deviations.

At the common 1,024-dimensional interface, the endpoint difference between Qwen3-0.6B and Qwen3-8B is approximately 0.020, substantially larger than the resampling variation. Increasing only the Qwen3-8B output width from 1,024 to 4,096 dimensions produces a smaller increase of approximately 0.005. The width effect is therefore detectable, but it is substantially smaller than the scale effect in this controlled comparison. This is consistent with the complete dimension-control results in Appendix E.

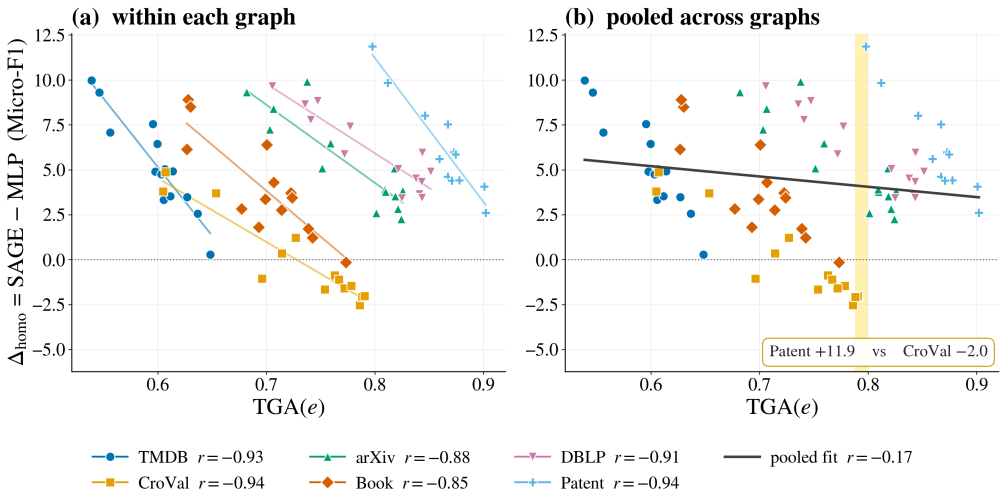


Figure 6: **TGA ranks homogeneous structural headroom within HTAG benchmarks.** Each point denotes a completed text-derived representation and $\Delta_{\text{homo}} = \text{MicroF1(SAGE)} - \text{MicroF1(MLP)}$. Within each HTAG, increasing TGA is strongly associated with decreasing homogeneous headroom. Pooling across HTAGs removes most of this relationship ($r = -0.17$), showing that absolute TGA values are not globally calibrated across graphs.

Within-HTAG Calibration TGA is designed to compare representations of the same graph rather than to provide a globally calibrated scale across different HTAGs. Across completed text-derived representations, the association between TGA and homogeneous structural headroom is strongly negative within every dataset, with Pearson correlations between -0.94 and -0.85. In contrast, pooling representation–graph pairs across datasets yields only $r = -0.17$.

The lack of cross-HTAG calibration is also visible in individual examples. Patent with TF–IDF has TGA 0.798 and $\Delta_{\text{homo}} = +11.87$, whereas CroVal with Qwen3-8B has a similar TGA of 0.788 but $\Delta_{\text{homo}} = -2.07$. TGA should therefore be used as a within-HTAG ranking signal. Its absolute value does not directly specify how much two different graphs will benefit from message passing.

D.6 Low-Rank Structure of the TGA Landscape

Although the exact label-only null is quantitatively rejected, its multiplicative form suggests that the observed TGA matrix may still be approximately separable into representation-side and graph-side factors.

For each dataset, define

$$M[e, m] = \text{TGA}(e, m) - \tfrac{1}{2}. \tag{17}$$

We fit the empirical form

$$M[e, m] = \gamma_{\mathcal{D}} \Big(A(e) - \tfrac{1}{2} \Big) \Big(h_m - h_0 \Big) + \varepsilon_{e,m}, \tag{18}$$

where $\gamma_{\mathcal{D}}$ is dataset-specific.

The leading rank-one component of M explains more than 99.5% of its energy on every dataset. The learned representation factor is strongly aligned with $A(e) - \frac{1}{2}$, with correlation $r_f \geq 0.944$ across datasets. The graph-side factor is similarly aligned with $h_m - h_0$, with $r_g \geq 0.967$ on five datasets and $r_g = 0.710$ on Patent. The fitted dataset-level scale lies in $\gamma_{\mathcal{D}} \in [2.70, 7.20]$.

These results distinguish the qualitative structure of the null from its absolute calibration. The coefficient-one form in equation 15 substantially underestimates the observed alignment, as reflected by the widespread upper-bound violations. At the same time, the near-rank-one structure indicates that the centered TGA values are well approximated by separable representation- and meta-path-specific factors.

E DIMENSION-CONTROLLED QWEN3 ANALYSIS

Qwen3 scaling changes both encoder capacity and native output dimensionality (Zhang et al., 2025). To separate these factors, we use its MRL interface and evaluate Qwen3 encoders at matched and native output dimensions. This yields two complementary controls. Encoder scale is varied at a common 1,024-dimensional interface, while output dimensionality is varied with Qwen3-8B held fixed.

Table 20 reports the complete TGA measurements. Table 21 reports the corresponding MLP, SAGE, and RSAGE Micro-F1 scores and structural gains.

Table 20: **TGA under dimension-controlled Qwen3 configurations.** Native denotes the encoder's default output width and MRL denotes a truncated Matryoshka output. The 1,024d rows isolate encoder scale, while the Qwen3-8B rows isolate output dimensionality.

Encoder	Output	Mode	TMDB	CroVal	arXiv	Book	DBLP	Patent
Fixed 1,024d interface								
Qwen3-0.6B	1024	Native	0.6114	0.7718	0.8090	0.7237	0.8212	0.9004
Qwen3-4B	1024	MRL	0.6265	0.7888	0.8204	0.7402	0.8399	0.9031
Qwen3-8B	1024	MRL	0.6328	0.7848	0.8189	0.7435	0.8355	0.9088
Higher-dimensional references								
Qwen3-4B	2560	Native	0.6268	0.7902	0.8209	0.7386	0.8408	0.9019
Qwen3-8B	2560	MRL	0.6344	0.7874	0.8228	0.7420	0.8406	0.9108
Qwen3-8B	4096	Native	0.6366	0.7881	0.8240	0.7423	0.8432	0.9148

Encoder scale at fixed width. At the common 1,024-dimensional interface, moving from Qwen3-0.6B to Qwen3-8B increases TGA on all six HTAGs and reduces Δ_{homo} on all six. Across datasets, the endpoint change is $+0.014$ in TGA and -1.50 Micro-F1 points in homogeneous headroom. Thus, the scale-dependent contraction observed in the main experiments persists when output dimensionality is held fixed.

Output width at fixed encoder scale. Holding Qwen3-8B fixed and varying only its output dimension produces a much smaller change in TGA. From 1,024 to 4,096 dimensions, mean TGA increases by only 0.004. Homogeneous headroom does not reproduce the scale-controlled contraction. Over the five HTAGs with completed message-passing runs at 4,096 dimensions, the mean endpoint change in Δ_{homo} is $+0.19$ points and the dataset-level changes have mixed signs. Patent is excluded from this comparison because the 4,096d message-passing runs exceed memory. These controls indicate that the reduction in homogeneous utility associated with encoder scaling cannot be explained by nominal output dimensionality alone.

Table 21: **Complete downstream results for dimension-controlled Qwen3.** Values are Micro-F1. Structural gains are $\Delta_{\text{homo}} = \text{SAGE} - \text{MLP}$ and $\Delta_{\text{rel}} = \text{RSAGE} - \text{SAGE}$. OOM denotes out-of-memory configurations.

Dataset	Configuration	MLP	SAGE	RSAGE	Δ_{homo}	Δ_{rel}
TMDB	0.6B@1024	77.12	80.66	84.58	+3.54	+3.92
	4B@1024	81.53	84.91	86.59	+3.38	+1.68
	8B@1024	84.62	85.36	88.74	+0.74	+3.38
	4B@2560	81.97	85.46	87.63	+3.49	+2.17
	8B@2560	84.49	86.20	88.56	+1.71	+2.36
	8B@4096	83.89	86.46	88.54	+2.57	+2.08
CroVal	0.6B@1024	90.19	88.60	90.65	−1.59	+2.05
	4B@1024	90.66	88.94	91.02	−1.72	+2.08
	8B@1024	90.63	88.91	91.15	−1.72	+2.24
	4B@2560	91.06	89.04	91.26	−2.02	+2.22
	8B@2560	91.09	89.16	91.25	−1.93	+2.09
	8B@4096	91.24	89.17	91.32	−2.07	+2.15
arXiv	0.6B@1024	80.09	84.01	84.33	+3.92	+0.32
	4B@1024	81.81	84.47	84.87	+2.66	+0.40
	8B@1024	82.04	84.35	85.09	+2.31	+0.74
	4B@2560	81.72	84.56	84.45	+2.84	−0.11
	8B@2560	82.11	84.43	84.53	+2.32	+0.10
	8B@4096	82.19	84.45	84.30	+2.26	−0.15
Book	0.6B@1024	79.26	82.72	83.65	+3.46	+0.93
	4B@1024	81.53	83.26	84.56	+1.73	+1.30
	8B@1024	81.99	83.43	84.72	+1.44	+1.29
	4B@2560	81.86	83.59	84.53	+1.73	+0.94
	8B@2560	82.68	83.72	84.73	+1.04	+1.01
	8B@4096	82.59	83.82	84.67	+1.23	+0.85
DBLP	0.6B@1024	70.99	76.05	77.01	+5.06	+0.96
	4B@1024	72.38	76.05	77.14	+3.67	+1.09
	8B@1024	72.40	76.14	76.90	+3.74	+0.76
	4B@2560	72.42	76.07	76.66	+3.65	+0.59
	8B@2560	72.69	76.19	76.35	+3.50	+0.16
	8B@4096	72.70	76.15	76.23	+3.45	+0.08
Patent	0.6B@1024	72.56	76.63	77.79	+4.07	+1.16
	4B@1024	74.17	77.12	78.32	+2.95	+1.20
	8B@1024	74.31	77.25	78.33	+2.94	+1.08
	4B@2560	74.53	77.14	76.38	+2.61	−0.76
	8B@2560	74.75	77.29	78.64	+2.54	+1.35
	8B@4096	74.90	OOM	OOM	—	—

F REPRESENTATION-INVARIANT STATISTICS

Classical homophily and related label–structure measures are defined on the pair (G, Y) alone, measuring how labels are distributed over adjacent nodes, which is fixed once the graph and the labels are. $(u, v) \in E$ denotes a target-node pair connected by a walk of length at most two, with relation types and intermediate node types discarded. Pair-based statistics use 30,000 sampled pairs with self-returns removed. The neighbor-distribution statistics (H_{node}, H_{agg}, CCNS, and h_S) use the endpoints of 12 sampled walks per anchor nodes, retaining returns to the anchor. They are invariant under any change of encoder and take a single value per HTAG, which leaves no within-HTAG variation.

(1) **Edge homophily ratio** (Zhu et al., 2020) is the fraction of connected pairs whose endpoints carry the same label:

$$H_{\text{edge}} = \frac{1}{|E|} \sum_{(u,v)\in E} \mathbf{1}[y_u = y_v]. \tag{19}$$

It weights each pair equally, so high-degree nodes dominate.

(2) **Node homophily** (Pei et al., 2020) averages the same agreement per node rather than per pair, so the two need not agree numerically:

$$H_{\text{node}} = \frac{1}{|V_T|} \sum_{v\in V_T} \frac{1}{d_v} \sum_{u\in\mathcal{N}(v)} \mathbf{1}[y_u = y_v]. \tag{20}$$

(3) **Class homophily** (Lim et al., 2021) credits each class only for agreement beyond what its size alone would produce:

$$h_c = \frac{\sum_{v\in V_c} \sum_{u\in\mathcal{N}(v)} \mathbf{1}[y_u = c]}{\sum_{v\in V_c} d_v}, \qquad H_{\text{class}} = \frac{1}{C-1} \sum_{c=1}^{C} \Big[h_c - \tfrac{|V_c|}{|V_T|}\Big]_+, \tag{21}$$

where $[\cdot]_+$ is the positive part.

(4) **Adjusted homophily** (Platonov et al., 2023) corrects H_{edge} for the degree-weighted class distribution, with D_c the total degree of class c:

$$H_{\text{adj}} = \frac{H_{\text{edge}} - \sum_c \big(D_c/2|E|\big)^2}{1 - \sum_c \big(D_c/2|E|\big)^2}, \tag{22}$$

so that randomly rewired edges give $H_{\text{adj}} = 0$.

(5) **Balanced homophily** (Platonov et al., 2023) averages the per-class conditional agreement with equal weight, so that a single large class cannot determine the value:

$$h_{\text{bal}} = \frac{1}{C} \sum_{k=1}^{C} h_k, \qquad h_k = \Pr(y_v = k \mid y_u = k), \tag{23}$$

where the conditional is taken over connected pairs. Classes absent from the sample are excluded from the average.

(6) **Adjusted balanced homophily** (Platonov et al., 2023) rescales h_{bal} against the uniform baseline $1/C$, keeping classes equally weighted while fixing the random-assignment value at 0:

$$h_{\text{bal}}^{\text{adj}} = \frac{C\, h_{\text{bal}} - 1}{C - 1} = \frac{\sum_k h_k - 1}{C - 1}. \tag{24}$$

(7) **Aggregation homophily** (Luan et al., 2022) asks whether post-aggregation similarity separates same-label from different-label nodes, using $S = \hat{A}Y(\hat{A}Y)^\top$, i.e. $S_{vu} = d(v)^\top d(u)$ for the neighbor label distribution $d(\cdot)$:

$$H_{\text{agg}} = \frac{1}{|V_T|} \sum_{v\in V_T} \mathbf{1}\Big[\operatorname*{mean}_{u:\, y_u = y_v} S_{vu} \;\geq\; \operatorname*{mean}_{u:\, y_u \neq y_v} S_{vu}\Big]. \tag{25}$$

Being an average of indicators, it saturates near its upper end on all six HTAGs.

Table 22: **Representation-invariant statistics across the six HTAGs.** Rows are the statistics (1)–(11) introduced in this section, columns are HTAGs. Each statistic is a function of the graph and the labels only, so a single value holds for all 15 node representations. Values are rounded to four decimal places. For Book, classes correspond to distinct label combinations: $C = 247$ for H_{class} and $C = 202$ for sampled balanced homophily and its adjusted version. For Patent, the corresponding sampled class count is $C = 118$.

	Statistic	TMDB	CroVal	arXiv	Book	DBLP	Patent
(1)	H_{edge}	0.5591	0.5753	0.5947	0.2115	0.6190	0.5853
(2)	H_{node}	0.6355	0.6257	0.5966	0.2377	0.6160	0.6240
(3)	H_{class}	0.3962	0.4008	0.3321	0.0427	0.5187	0.3483
(4)	H_{adj}	0.4095	0.4139	0.5411	0.1789	0.5434	0.5548
(5)	h_{bal}	0.5472	0.5007	0.3488	0.0568	0.5722	0.3597
(6)	$h_{\text{bal}}^{\text{adj}}$	0.3962	0.4008	0.3321	0.0521	0.5187	0.3542
(7)	H_{agg}	0.9500	0.8692	0.8616	0.8480	0.8561	0.9324
(8)	CCNS gap	0.4436	0.3449	0.2623	0.1332	0.4046	0.4237
(9)	h_{mod}	0.3058	0.2999	0.4779	0.1718	0.4534	0.5168
(10)	h_S	0.6635	0.6321	0.9019	0.9848	0.6761	0.9742
(11)	LI	0.2584	0.1899	0.3792	0.1845	0.3271	0.4709

(8) CCNS gap (Ma et al., 2021). Cross-class neighborhood similarity compares the neighbor label distributions of two classes,

$$s(c,c') = \begin{cases} \dfrac{1}{|V_c|(|V_c|-1)} \displaystyle\sum_{\substack{i,j \in V_c \\ i \neq j}} \cos\big(d(i), d(j)\big), & c = c', \\ \dfrac{1}{|V_c||V_{c'}|} \displaystyle\sum_{i \in V_c} \sum_{j \in V_{c'}} \cos\big(d(i), d(j)\big), & c \neq c'. \end{cases} \tag{26}$$

Self-pairs are excluded from the diagonal term ($c = c'$), and classes with fewer than two sampled nodes are excluded from the diagonal average. We summarize the resulting $C \times C$ matrix by the gap between its diagonal and off-diagonal means.

(9) Modularity (Platonov et al., 2023) measures the excess of within-class edges over the configuration-model expectation, with $\bar{p}(k) = D_k/2|E|$ the degree-weighted class share:

$$h_{\text{mod}} = H_{\text{edge}} - \sum_k \bar{p}(k)^2, \qquad \text{so that} \qquad H_{\text{adj}} = \frac{h_{\text{mod}}}{1 - \sum_k \bar{p}(k)^2}. \tag{27}$$

(10) Tri-Hom structural homophily h_S (Zheng et al., 2024) measures how concentrated the neighbor label distributions are within each class:

$$h_S = \frac{1}{C} \sum_{c=1}^{C} \Big[1 - \frac{\sigma\big(\{d(u) : y_u = c\}\big)}{\sigma_{\max}} \Big], \tag{28}$$

where $\sigma(\cdot)$ is the mean of the per-coordinate standard deviations and $\sigma_{\max} = 0.5$. The average is taken only over classes with at least two sampled nodes, with C denoting the number of such classes.

(11) Label informativeness (Platonov et al., 2023) is the mutual information between the labels at the two ends of a random connected pair, normalized by the endpoint label entropy:

$$\text{LI} = \frac{I(Y_u; Y_v)}{H(Y_{\text{endpoint}})}. \tag{29}$$

G REPRESENTATION-DEPENDENT STATISTICS

We compare TGA with four types of statistics and two label-informed references, each capturing a different aspect of an HTAG. *Embedding geometry* statistics depend only on the node-representation matrix $X^{(e)} \in \mathbb{R}^{N \times d_e}$ and characterize the geometry of the text-representation space independently of graph connectivity. *Feature homophily* measures whether connected nodes have similar features. *Label distinguishability* uses labels to measure how well a representation separates classes, reported before and after one aggregation step. *Similarity alignment* statistics use graph structure and $X^{(e)}$ jointly without labels. They are therefore the closest existing points of comparison for TGA, but they summarize similarity through raw or mean cosine values rather than through the rank comparison used by TGA. For homogeneous-graph baselines, we use an untyped target-node projection A_U, in which two target nodes are connected if a non-self walk of length at most two connects them, discarding relation types and intermediate node types. On the multi-label **Book** dataset, statistics that condition on $y_u = y_v$ treat nodes with identical label sets as belonging to the same class. The raw values in Table 23 are not compared across statistics; Section 7.3 instead evaluates each statistic through its within-HTAG rank association with Δ_{homo} and Δ_{rel}.

Embedding geometry. Let $\mathcal{S}$ be a uniform sample of at most 20,000 target nodes, let $n = |\mathcal{S}|$, and define $\tilde{x}_i = x_i / \max(\|x_i\|_2, \varepsilon)$.

(1) **Anisotropy** (Ethayarajh, 2019) is the mean cosine similarity over all distinct pairs in the node sample. We compute it in its equivalent closed form, rather than by sampling pairs:

$$\text{Aniso}(X^{(e)}) = \frac{\left\|\sum_{i \in \mathcal{S}} \tilde{x}_i\right\|_2^2 - n}{n(n-1)}. \tag{30}$$

Its expectation is 0 for isotropic random representations and approaches 1 when all vectors share a common direction.

(2) **IsoScore** (Rudman et al., 2022) measures how uniformly a representation utilizes the ambient dimensions. Let $\lambda = (\lambda_1, \ldots, \lambda_d)$ be the eigenvalues of the covariance matrix of the normalized sampled representations $\{\tilde{x}_i\}_{i \in \mathcal{S}}$, i.e., the variance vector of the PCA-reoriented sample, and let $\hat{\lambda} = \sqrt{d}\,\lambda / \|\lambda\|$. With the isotropy defect $\delta = \|\hat{\lambda} - \mathbf{1}\| / \sqrt{2(d - \sqrt{d})}$, the score is

$$\text{IsoScore}(X^{(e)}) = \frac{\left(d - \delta^2(d - \sqrt{d})\right)^2 - d}{d(d-1)}. \tag{31}$$

It equals 1 when variance is spread uniformly over all d directions and 0 when it concentrates in a single direction, and approximates the fraction of ambient dimensions uniformly utilized.

(3) **RankMe** (Garrido et al., 2023) is a smooth approximation to the rank of $X^{(e)}$. Let $\sigma_1 \geq \cdots \geq \sigma_r$ be the singular values of the uncentered matrix whose rows are the normalized sample vectors $\tilde{x}_i$, $i \in \mathcal{S}$, where $r = \min(n, d)$, and $\epsilon = 10^{-7}$ is used as a small constant for numerical stability.

$$p_k = \frac{\sigma_k + \epsilon}{\sum_{\ell=1}^{r}(\sigma_\ell + \epsilon)}, \qquad \text{RankMe}(X^{(e)}) = \exp\left(-\sum_{k=1}^{r} p_k \log p_k\right) \tag{32}$$

It ranges from 1 to r and was proposed as a label-free proxy for linear-probe accuracy in self-supervised representations.

Feature homophily.

(4) **Tri-Hom feature homophily** h_F^* (Zheng et al., 2024) models the observed features as a linear diffusion, over the graph, of latent features $X^{(e,0)}$ that are independent of graph structure. For each feature dimension m,

$$X^{(e)}_{:,m} = \left(I - \frac{h_{F,m}}{\rho(A_U)} A_U\right)^{-1} X^{(e,0)}_{:,m}, \qquad h_{F,m} \in (-1, 1), \tag{33}$$

where $\rho(A_U)$ is the spectral radius of A_U, so the bound keeps the inverse well defined. Since $h_{F,m}$ and $X^{(e,0)}$ are both unknown, equation 33 alone does not identify $h_{F,m}$. Following

Zheng et al. (2024), we write $X^{(e,0)}_{:,m}(h) = \big(I - \frac{h}{\rho(A_U)} A_U\big) X^{(e)}_{:,m}$ for the latent features implied by a candidate h, and pick the value that makes them most similar within each class:

$$h^*_{F,m} = \operatorname*{arg\,min}_{h\in(-1,1)} \sum_{\substack{u,v\in V_T \\ y_u = y_v}} \big(X^{(e,0)}_{u,m}(h) - X^{(e,0)}_{v,m}(h)\big)^2, \qquad h^*_F = \frac{1}{d_e}\sum_{m=1}^{d_e} h^*_{F,m}. \tag{34}$$

We compute h^*_F on a 3,000-node snowball subgraph of A_U.

Label distinguishability.

(5) Distance ratio J (Luan et al., 2023) compares within-label and between-label squared distances of a representation H, with lower values indicating better separation:

$$J(H) = \frac{\mathbb{E}\big[\|h_u - h_v\|_2^2 \mid y_u = y_v\big]}{\mathbb{E}[\|h_u - h_v\|_2^2 \mid y_u \neq y_v]}. \tag{35}$$

Both expectations use 200,000 independently sampled non-self pairs, counting identical label vectors as one configuration in the multi-label datasets. We report J on $X^{(e)}$ and on $\widetilde{A}_U X^{(e)}$ using the same anchor nodes, where the latter averages sampled walk-endpoint features, including returns to the anchor.

Similarity alignment. Let $\widehat{\mathcal{E}}_U = \{(u_k, v_k)\}_{k=1}^{K_E}$ denote $K_E = 30{,}000$ target-node pairs sampled from the untyped projection, and let $\widehat{\mathcal{R}} = \{(r_k, s_k)\}_{k=1}^{K_R}$ denote independently sampled non-self target-node pairs, where $K_R \approx K_E$.

(6) Edge similarity. The feature-based analogue of edge homophily replaces label equality with cosine similarity, estimating the attribute assortativity in the graph:

$$H_{\mathrm{GE}}(X^{(e)}) = \frac{1}{K_E}\sum_{k=1}^{K_E} \cos(x_{u_k}, x_{v_k}). \tag{36}$$

Its absolute value is strongly affected by anisotropy: when the feature space is highly anisotropic, nearly all node pairs appear similar.

(7) Mean gap (untyped). We compare projection-connected pairs with independently sampled random pairs:

$$\Delta^U_{\cos}(X^{(e)}) = \frac{1}{K_E}\sum_{k=1}^{K_E} \cos(x_{u_k}, x_{v_k}) - \frac{1}{K_R}\sum_{k=1}^{K_R} \cos(x_{r_k}, x_{s_k}). \tag{37}$$

Thus, a value near 0 indicates no excess feature similarity over the random-pair reference.

(8) Mean gap (typed). For each meta-path m, we use the same connected and random pair samples $\mathcal{P}_m$ and $\mathcal{R}_m$ as TGA, with $K_m = |\mathcal{P}_m| = |\mathcal{R}_m|$.

$$\Delta^T_{\cos}(X^{(e)}) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \Bigg[\frac{1}{K_m} \sum_{(u,v)\in\mathcal{P}_m} \cos(x^{(e)}_u, x^{(e)}_v) - \frac{1}{K_m} \sum_{(r,s)\in\mathcal{R}_m} \cos(x^{(e)}_r, x^{(e)}_s) \Bigg]. \tag{38}$$

Label-informed references. The following two quantities use labels and are reported as reference points rather than as competing statistics.

(9) Class separation $A(e)$ applies the comparison in Eq. 2 to label-defined pairs instead of meta-path-connected ones, reading the representation without the graph. With $U(\cdot,\cdot)$ the tie-corrected Mann–Whitney statistic,

$$A(e) = U\big(\{\cos(x_u, x_v) : y_u = y_v\},\ \{\cos(x_a, x_b) : y_a \neq y_b\}\big), \tag{39}$$

so that $A(e) = 0.5$ under no class structure.

(10) MLP Micro-F1 (oracle). The test Micro-F1 of the text-only MLP, taken at its best-validation epoch, summarizes what the representation alone resolves.

Table 23: **Baseline statistics across all datasets and node representations.** Rows are the statistics (1)–(10) introduced above, blocked by dataset; columns are the 15 node representations, grouped into control, static, PLM, retrieval, and LLM-based families. All values are rounded to four decimal places, except RankMe and the MLP test Micro-F1 (in %), which are rounded to two.

	Control	Static		PLM						Retrieval		LLM-based			
Statistic	*Random*	TF-IDF	GloVe	DistilBERT	SimCSE	MPNet	SPECTER2	MiniLM	Sentence-T5	BGE-large	BGE-M3	Qwen3-0.6B	Qwen3-4B	Qwen3-8B	NV-Embed-v2
TMDB															
(1) Aniso	0.0000	0.0161	0.5456	0.8043	0.5518	0.2260	0.8920	0.1524	0.6728	0.4203	0.4459	0.3647	0.3536	0.3307	0.1661
(2) IsoScore	0.9511	0.6478	0.1867	0.0496	0.0901	0.1142	0.0841	0.2763	0.1213	0.1219	0.1597	0.1116	0.0572	0.0356	0.0485
(3) RankMe	381.51	970.26	198.69	312.88	372.22	445.95	257.07	289.33	397.28	586.77	588.07	643.49	1405.15	2086.47	2364.00
(4) h_F^*	-0.0166	0.0155	0.1224	0.1053	0.1191	0.1501	0.0395	0.1342	0.0684	0.1053	0.0927	0.1066	0.1128	0.1231	0.1601
(5) $J(X)$	0.9998	0.9936	0.9249	0.8756	0.9223	0.9206	0.9523	0.9429	0.9201	0.9368	0.9526	0.9329	0.9356	0.9289	0.9372
$J(\widetilde{A}_U X)$	1.0004	0.9904	0.8795	0.8083	0.8778	0.8778	0.9260	0.9119	0.8786	0.9018	0.9202	0.8915	0.8925	0.8865	0.9020
(6) H_{GE}	-0.0002	0.0325	0.5739	0.8242	0.5853	0.2837	0.8972	0.2037	0.6971	0.4564	0.4744	0.4074	0.3988	0.3798	0.2227
(7) $\Delta_{\cos}^U$	-0.0006	0.0162	0.0289	0.0200	0.0334	0.0572	0.0050	0.0505	0.0242	0.0362	0.0279	0.0426	0.0450	0.0490	0.0561
(8) $\Delta_{\cos}^T$	0.0001	0.0120	0.0261	0.0183	0.0287	0.0461	0.0039	0.0395	0.0206	0.0308	0.0235	0.0340	0.0368	0.0395	0.0478
(9) $A(e)$	0.5027	0.5309	0.5712	0.6352	0.6431	0.6502	0.5567	0.6393	0.6609	0.6364	0.6203	0.6658	0.6725	0.7021	0.6878
(10) MLP	25.22	66.30	70.88	74.09	75.31	77.79	61.77	72.19	80.53	79.89	77.21	77.12	81.97	83.89	87.63
CroVal															
(1) Aniso	-0.0000	0.0502	0.7243	0.8775	0.7589	0.0675	0.8904	0.0739	0.7739	0.5387	0.4519	0.3051	0.3050	0.2658	0.1815
(2) IsoScore	0.9811	0.4619	0.0974	0.0134	0.0661	0.0929	0.0440	0.2004	0.0898	0.0990	0.1432	0.0827	0.0400	0.0292	0.0314
(3) RankMe	383.07	943.19	164.96	226.15	303.38	474.04	250.90	289.23	361.15	568.69	592.29	641.25	1415.83	2169.07	2332.22
(4) h_F^*	-0.0103	0.1673	0.1810	0.1304	0.1533	0.4756	0.1869	0.4475	0.1332	0.2233	0.2114	0.3430	0.3414	0.3645	0.3889
(5) $J(X)$	1.0001	0.9735	0.9278	0.9428	0.8731	0.8918	0.7710	0.9027	0.8566	0.8828	0.9209	0.8610	0.8716	0.8936	0.8898
$J(\widetilde{A}_U X)$	0.9799	0.9524	0.9080	0.9096	0.8676	0.8911	0.7741	0.8955	0.8517	0.8747	0.9131	0.8632	0.8732	0.8900	0.8879
(6) H_{GE}	0.0001	0.1240	0.7609	0.8975	0.7942	0.2461	0.9179	0.2383	0.8128	0.6159	0.5205	0.4368	0.4389	0.3988	0.3258
(7) $\Delta_{\cos}^U$	0.0008	0.0742	0.0369	0.0204	0.0359	0.1792	0.0278	0.1651	0.0387	0.0774	0.0690	0.1322	0.1348	0.1337	0.1447
(8) $\Delta_{\cos}^T$	-0.0000	0.0717	0.0364	0.0180	0.0341	0.1800	0.0287	0.1631	0.0363	0.0762	0.0707	0.1334	0.1409	0.1393	0.1426
(9) $A(e)$	0.4940	0.6829	0.5769	0.5495	0.6354	0.7838	0.7614	0.7684	0.7118	0.7950	0.7095	0.8211	0.8219	0.8078	0.8311
(10) MLP	46.65	85.45	80.76	77.81	82.11	89.72	88.79	88.62	88.81	89.03	86.02	90.19	91.06	91.24	91.84
arXiv															
(1) Aniso	0.0000	0.0556	0.7986	0.9067	0.7440	0.1358	0.8074	0.1391	0.7878	0.6051	0.4809	0.2800	0.3698	0.3301	0.2228
(2) IsoScore	0.9810	0.5721	0.1243	0.0417	0.0625	0.0758	0.0432	0.1498	0.1064	0.0646	0.1438	0.0941	0.0516	0.0324	0.0336
(3) RankMe	383.07	970.69	156.26	222.11	295.79	497.18	261.19	289.57	359.01	553.29	582.13	664.77	1462.95	2204.61	2310.68
(4) h_F^*	-0.0017	0.1948	0.2624	0.1369	0.2441	0.6503	0.4311	0.6010	0.2197	0.3520	0.3429	0.5673	0.5306	0.5612	0.6169
(5) $J(X)$	0.9995	0.9577	0.8260	0.7660	0.8031	0.7784	0.7620	0.7769	0.8019	0.7914	0.8893	0.8378	0.8638	0.8546	0.8340
$J(\widetilde{A}_U X)$	0.9012	0.8419	0.7018	0.6556	0.6835	0.6410	0.6265	0.6409	0.6760	0.6462	0.7601	0.6936	0.7071	0.6967	0.7117
(6) H_{GE}	-0.0002	0.1772	0.8570	0.9359	0.8144	0.5039	0.8969	0.4749	0.8535	0.7503	0.6015	0.5237	0.5938	0.5692	0.4743
(7) $\Delta_{\cos}^U$	-0.0002	0.1218	0.0587	0.0290	0.0701	0.3682	0.0896	0.3361	0.0658	0.1450	0.1204	0.2440	0.2241	0.2395	0.2520
(8) $\Delta_{\cos}^T$	-0.0000	0.0709	0.0386	0.0201	0.0472	0.2384	0.0593	0.2207	0.0425	0.0956	0.0735	0.1524	0.1369	0.1466	0.1569
(9) $A(e)$	0.4965	0.6893	0.6302	0.6624	0.6725	0.8498	0.8559	0.8365	0.7237	0.8255	0.7223	0.8249	0.8203	0.8322	0.8012
(10) MLP	35.49	73.22	72.27	73.74	73.74	80.50	80.49	78.87	78.07	80.10	76.56	80.09	81.72	82.19	81.82

continued on next page

Book

(1)	Aniso	0.0000	0.0378	0.6783	0.8560	0.6312	0.1773	0.8910	0.1643	0.7490	0.4889	0.4217	0.2751	0.2529	0.2988	0.1488
(2)	IsoScore	0.9812	0.7008	0.1177	0.0483	0.0818	0.1219	0.1012	0.2894	0.1138	0.1254	0.1741	0.1375	0.0700	0.0401	0.0648
(3)	RankMe	383.08	982.91	178.81	278.91	354.41	466.04	282.43	291.44	384.80	604.35	618.75	692.08	1568.25	2305.86	2620.18
(4)	h_F^*	-0.0215	0.1038	0.2881	0.2119	0.2879	0.3489	0.1280	0.3325	0.1766	0.2301	0.2334	0.3041	0.3225	0.3065	0.3766
(5)	$J(X)$	1.0001	0.9855	0.9133	0.8711	0.9098	0.9280	0.9982	0.9490	0.9028	0.9264	0.9545	0.9483	0.9467	0.9411	0.9476
	$J(\widetilde{A}_U X)$	1.0232	0.9878	0.9171	0.8291	0.8736	0.9020	0.9854	0.9283	0.8595	0.8909	0.9292	0.9197	0.9100	0.9025	0.9084
(6)	H_{GE}	-0.0000	0.0734	0.7319	0.8871	0.6877	0.2953	0.9020	0.2566	0.7901	0.5548	0.4767	0.3596	0.3434	0.3912	0.2545
(7)	$\Delta_{\cos}^U$	-0.0003	0.0360	0.0572	0.0315	0.0569	0.1203	0.0114	0.0939	0.0414	0.0666	0.0560	0.0851	0.0915	0.0936	0.1064
(8)	$\Delta_{\cos}^T$	0.0001	0.0422	0.0558	0.0305	0.0583	0.1220	0.0120	0.1008	0.0418	0.0688	0.0611	0.0914	0.0971	0.0987	0.1133
(9)	$A(e)$	0.5006	0.5778	0.5846	0.6287	0.6476	0.6701	0.5754	0.6436	0.6394	0.6578	0.6505	0.6785	0.6914	0.6928	0.7013
(10)	MLP	21.26	73.31	75.06	79.15	78.69	78.55	71.66	75.07	81.01	79.92	78.06	79.26	81.86	82.59	84.10

DBLP

(1)	Aniso	-0.0000	0.0479	0.7687	0.8931	0.7274	0.0944	0.8033	0.0904	0.7744	0.5870	0.4682	0.2588	0.3594	0.3266	0.1946
(2)	IsoScore	0.9811	0.6149	0.1106	0.0345	0.0580	0.0991	0.0605	0.1905	0.1145	0.0746	0.1501	0.1151	0.0591	0.0376	0.0446
(3)	RankMe	383.07	973.17	161.80	228.15	299.46	504.87	288.17	295.03	363.25	563.17	583.36	674.88	1480.71	2220.82	2347.82
(4)	h_F^*	0.0137	0.2782	0.3030	0.1726	0.2815	0.7268	0.3995	0.6772	0.2465	0.4090	0.4188	0.5679	0.5378	0.5897	0.6483
(5)	$J(X)$	1.0000	0.9844	0.8993	0.8599	0.9019	0.9125	0.8904	0.9202	0.9178	0.9087	0.9531	0.9336	0.9358	0.9351	0.9374
	$J(\widetilde{A}_U X)$	1.0131	0.9802	0.8591	0.8160	0.8707	0.8805	0.8573	0.8887	0.8896	0.8704	0.9328	0.9096	0.9041	0.9042	0.9178
(6)	H_{GE}	-0.0002	0.1777	0.8439	0.9310	0.8097	0.4859	0.8907	0.4514	0.8477	0.7471	0.5980	0.5106	0.5964	0.5801	0.4568
(7)	$\Delta_{\cos}^U$	-0.0005	0.1294	0.0746	0.0379	0.0823	0.3911	0.0872	0.3600	0.0731	0.1596	0.1296	0.2522	0.2372	0.2532	0.2618
(8)	$\Delta_{\cos}^T$	-0.0002	0.0722	0.0509	0.0278	0.0557	0.2457	0.0583	0.2297	0.0465	0.1035	0.0761	0.1528	0.1408	0.1501	0.1568
(9)	$A(e)$	0.4976	0.5980	0.5904	0.6231	0.6172	0.7225	0.7155	0.7094	0.6367	0.7133	0.6357	0.6864	0.7005	0.7032	0.6834
(10)	MLP	24.82	66.99	65.35	67.27	66.62	71.05	71.60	69.86	69.50	71.46	68.11	70.99	72.42	72.70	72.83

Patent

(1)	Aniso	0.0000	0.0311	0.5859	0.8350	0.6441	0.2246	0.8311	0.1646	0.7442	0.4980	0.4814	0.2344	0.3232	0.2896	0.1558
(2)	IsoScore	0.9813	0.5213	0.0924	0.0314	0.0472	0.1667	0.0710	0.2664	0.0966	0.1357	0.1060	0.1311	0.0702	0.0419	0.0510
(3)	RankMe	383.08	963.22	180.61	257.19	308.71	497.61	290.09	294.60	357.86	595.47	555.78	685.65	1521.76	2249.63	2383.13
(4)	h_F^*	-0.0202	0.2498	0.2567	0.1307	0.2321	0.4214	0.2329	0.4278	0.1734	0.2804	0.2780	0.3700	0.3497	0.3765	0.4566
(5)	$J(X)$	0.9998	0.9684	0.7740	0.7899	0.8306	0.9133	0.8542	0.9031	0.8686	0.9107	0.9031	0.9134	0.9024	0.9047	0.9209
	$J(\widetilde{A}_U X)$	0.9527	0.8803	0.6268	0.6339	0.6752	0.7990	0.7210	0.7827	0.7281	0.7915	0.7688	0.7825	0.7840	0.7771	0.7966
(6)	H_{GE}	0.0005	0.1914	0.7421	0.9034	0.7786	0.4588	0.8937	0.4195	0.8359	0.6398	0.6243	0.4594	0.5258	0.5023	0.3977
(7)	$\Delta_{\cos}^U$	0.0012	0.1608	0.1565	0.0683	0.1346	0.2342	0.0628	0.2553	0.0923	0.1420	0.1434	0.2256	0.2034	0.2136	0.2427
(8)	$\Delta_{\cos}^T$	0.0001	0.1368	0.1431	0.0639	0.1247	0.2094	0.0577	0.2296	0.0840	0.1271	0.1281	0.2027	0.1828	0.1912	0.2164
(9)	$A(e)$	0.4975	0.6667	0.7149	0.7342	0.7601	0.7321	0.7557	0.7403	0.7512	0.7478	0.7400	0.7797	0.7790	0.7983	0.7720
(10)	MLP	13.26	61.19	63.99	66.38	67.21	71.76	69.56	69.64	71.96	72.01	69.97	72.56	74.53	74.90	74.87

H Correlation Robustness Across Coefficients

We examine whether the association between TGA and residual homogeneous headroom depends on the choice of correlation coefficient. Table 24 reports mean absolute within-HTAG correlations for the evaluated label-free statistics. TGA has the strongest mean association under Pearson, Spearman, and Kendall coefficients. Table 25 reports the corresponding per-HTAG results for TGA. The association is negative and statistically significant at the 0.01 level in every HTAG under all three coefficients. The association also remains negative in all 24 leave-one-representation-family-out analyses over the four families in Table 23, with the weakest Spearman's ρ at -0.430.

Table 24: **Correlation robustness across evaluated statistics.** Mean absolute within-HTAG correlations across the six HTAGs. Representation-invariant statistics have undefined within-HTAG correlations.

		Δ_{homo}			Δ_{rel}		
Category	Statistic	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
Structure	homophily & related	constant in e					
Graph–label dependence	label informativeness	constant in e					
Geometry	anisotropy	0.244	0.170	0.142	0.372	0.327	0.240
	IsoScore	0.476	0.456	0.395	0.374	0.362	0.253
	RankMe	0.529	0.596	0.489	0.490	0.540	0.425
Label distinguishability	dist. ratio J	0.348	0.323	0.239	**0.525**	**0.621**	**0.479**
Feature homophily	Tri-Hom h_F^*	0.626	0.536	0.381	0.329	0.287	0.181
Similarity alignment	edge sim.	0.172	0.177	0.142	0.401	0.336	0.239
	mean gap (U)	0.551	0.496	0.367	0.279	0.269	0.204
	mean gap (T)	0.556	0.499	0.376	0.250	0.256	0.188
Rank alignment	**TGA**	**0.909**	**0.800**	**0.654**	0.390	0.310	0.254
Label-informed reference	class sep. $A(e)$	0.884	0.764	0.624	0.339	0.289	0.218
	MLP	0.984	0.965	0.894	0.454	0.510	0.417

Bold marks the strongest associations excluding MLP. U denotes untyped target-node projection and T typed target-returning meta-path connectivity. The distance ratio is computed on aggregated features, $J(\widetilde{A}_U X^{(e)})$.

Table 25: **TGA and residual homogeneous headroom.** Within-HTAG correlations with Δ_{homo}.

Dataset	Pearson		Spearman			Kendall	
	r	p	ρ	p	95% CI	τ	p
TMDB	-0.926	<0.0001	-0.873	0.0002	$[-0.995, -0.530]$	-0.758	0.0002
CroVal	-0.940	<0.0001	-0.895	0.0001	$[-0.973, -0.642]$	-0.714	0.0003
arXiv	-0.881	0.0001	-0.785	0.0015	$[-0.948, -0.399]$	-0.604	0.0021
Book	-0.853	0.0002	-0.705	0.0056	$[-0.941, -0.204]$	-0.560	0.0046
DBLP	-0.911	0.0001	-0.719	0.0051	$[-0.951, -0.197]$	-0.560	0.0042
Patent	-0.939	<0.0001	-0.825	0.0015	$[-1.000, -0.374]$	-0.727	0.0006

p-values use unadjusted two-sided tests with 20,000 permutations. Correlations use 14 text-derived representations per HTAG, except Patent ($n = 12$).

I GEOMETRY OF NODE REPRESENTATIONS

We complement the representation sweep with a qualitative analysis of all 15 benchmark representations across the six HTAGs. Figures 7–12 show their first three principal components for 6,000 sampled target nodes per dataset. Colors distinguish up to six of the most frequent classes, with four classes on TMDB. Each panel reports anisotropy in the original embedding space and the variance explained by PC1–3. Representations follow a common order based on mean anisotropy across the six datasets, so anisotropy need not increase monotonically within an individual dataset.

Across the six HTAGs, projected geometry and class overlap vary across representations, but lower anisotropy does not consistently correspond to clearer class separation. These observations remain qualitative, since PC1–3 capture only part of the total variance. The projections alone do not establish full-dimensional class separability or the remaining benefit of message passing.

These visualizations complement the quantitative comparison in Section 7.3. Together, they motivate examining node representations jointly with graph connectivity when assessing residual homogeneous headroom.

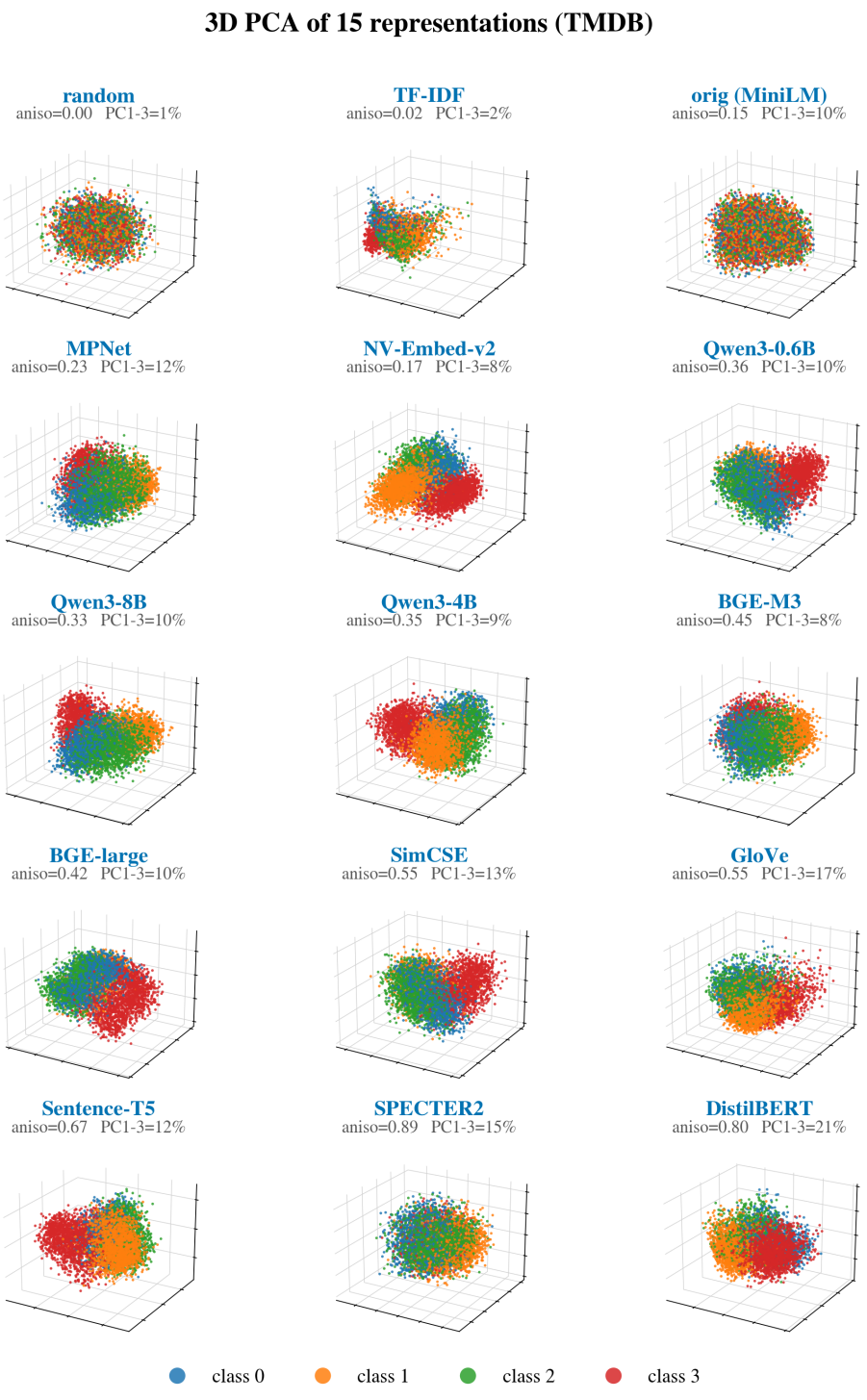


Figure 7: First three principal components of all 15 representations on TMDB, for 6,000 sampled target nodes colored by the most frequent classes (up to six; TMDB has four classes in total). Each panel is annotated with its anisotropy computed over all target nodes and the share of total variance captured by PC1–3. Panels are ordered by anisotropy averaged over the six datasets, ascending, identically in all six figures, so a given grid position always refers to the same representation; per-dataset anisotropy therefore increases only approximately along the reading order, and the annotated values give the exact local ordering. Class separability does not follow that order. The same layout is used in Figures 8–12.

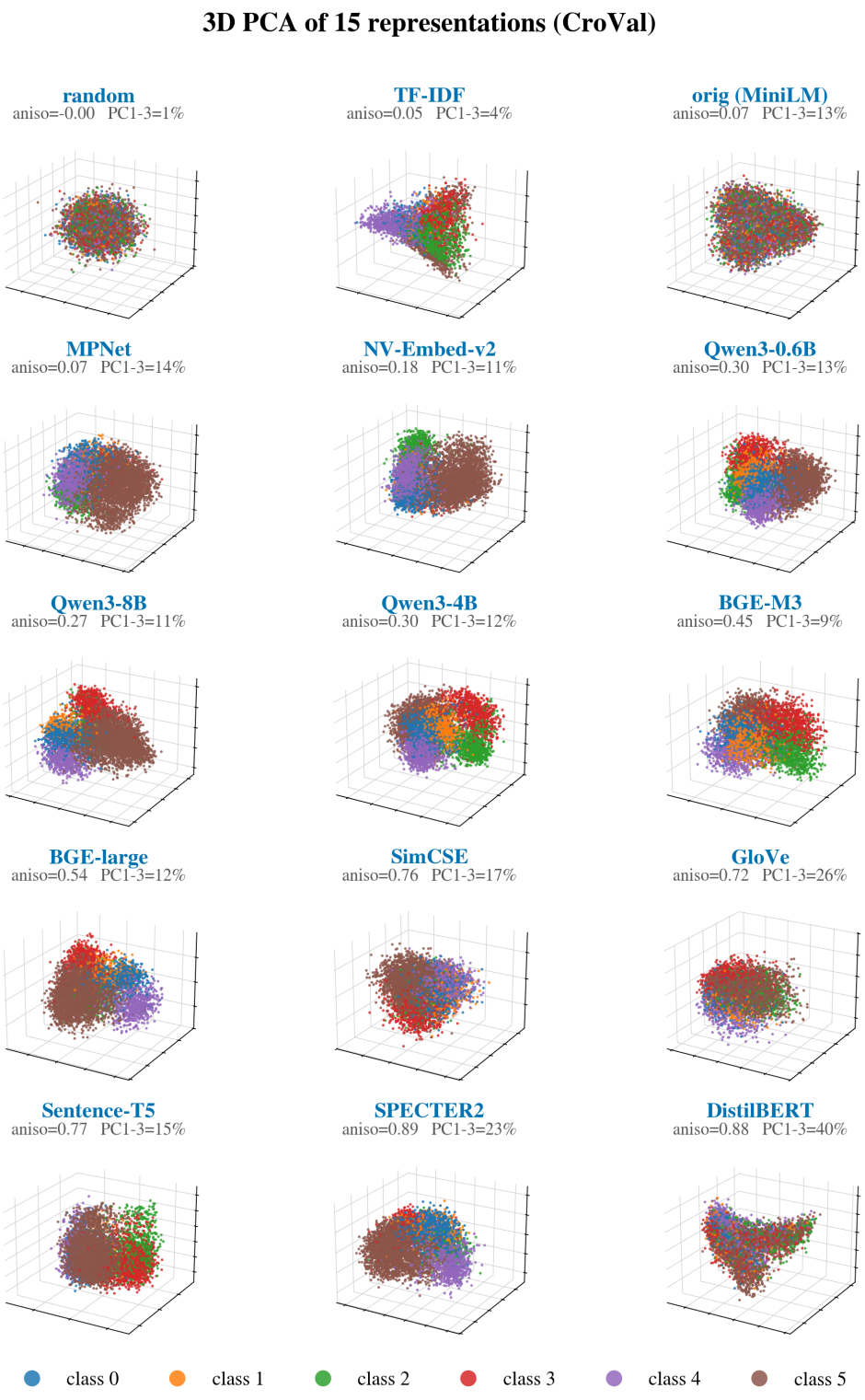


Figure 8: First three principal components of all 15 representations on CroVal. Layout, sampling, and annotations follow Figure 7.

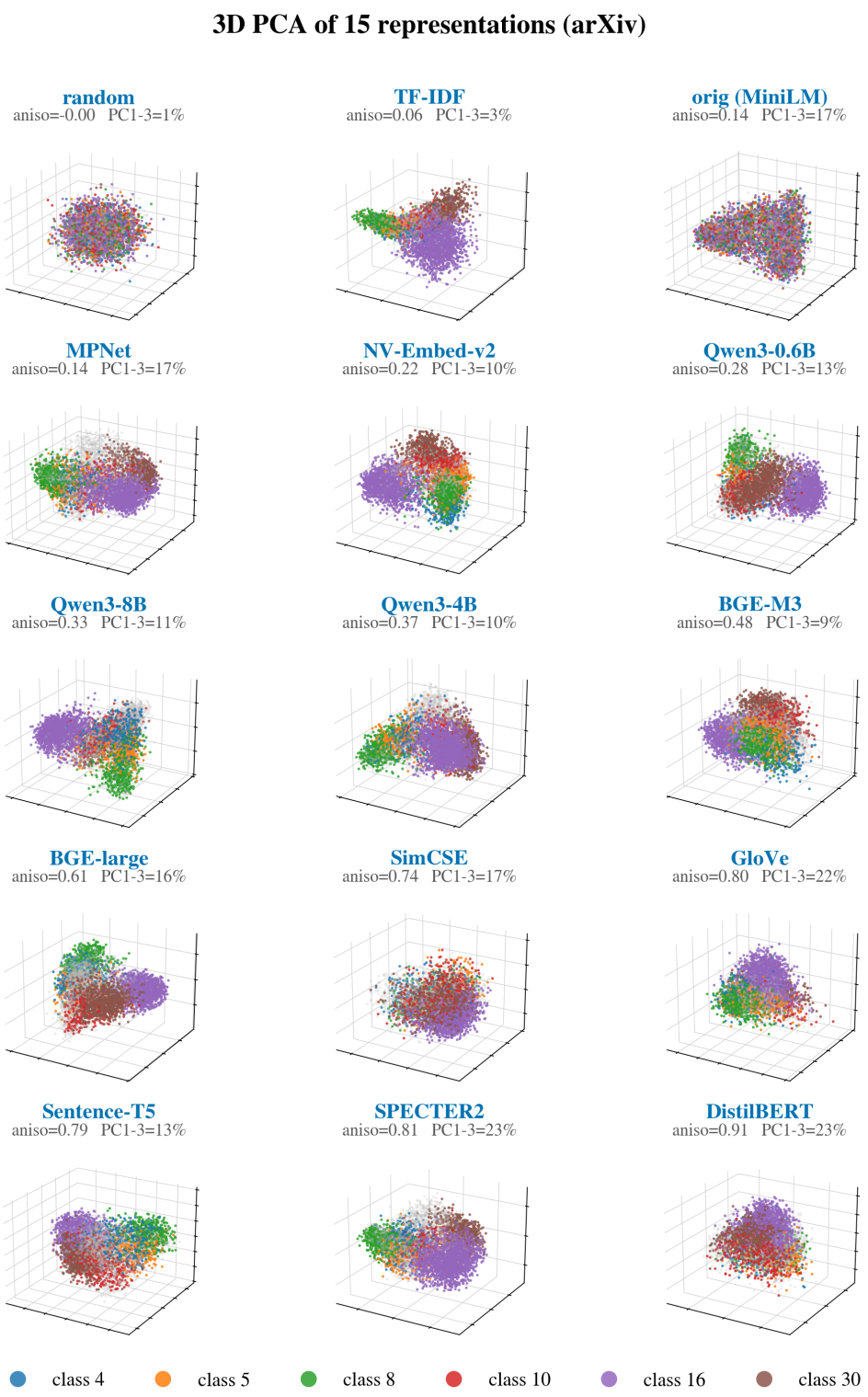


Figure 9: First three principal components of all 15 representations on arXiv. Layout, sampling, and annotations follow Figure 7.

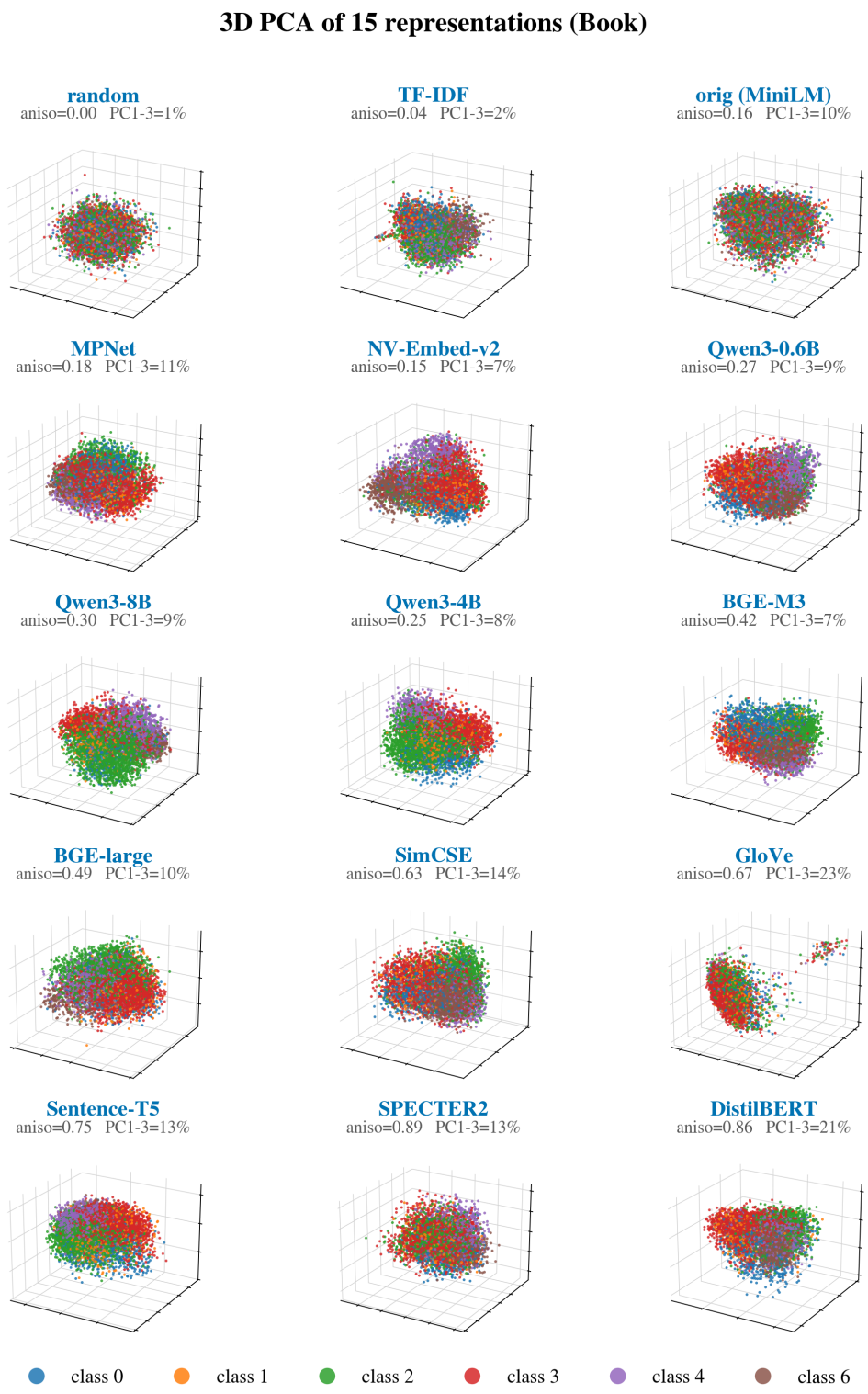


Figure 10: First three principal components of all 15 representations on Book. Layout, sampling, and annotations follow Figure 7.

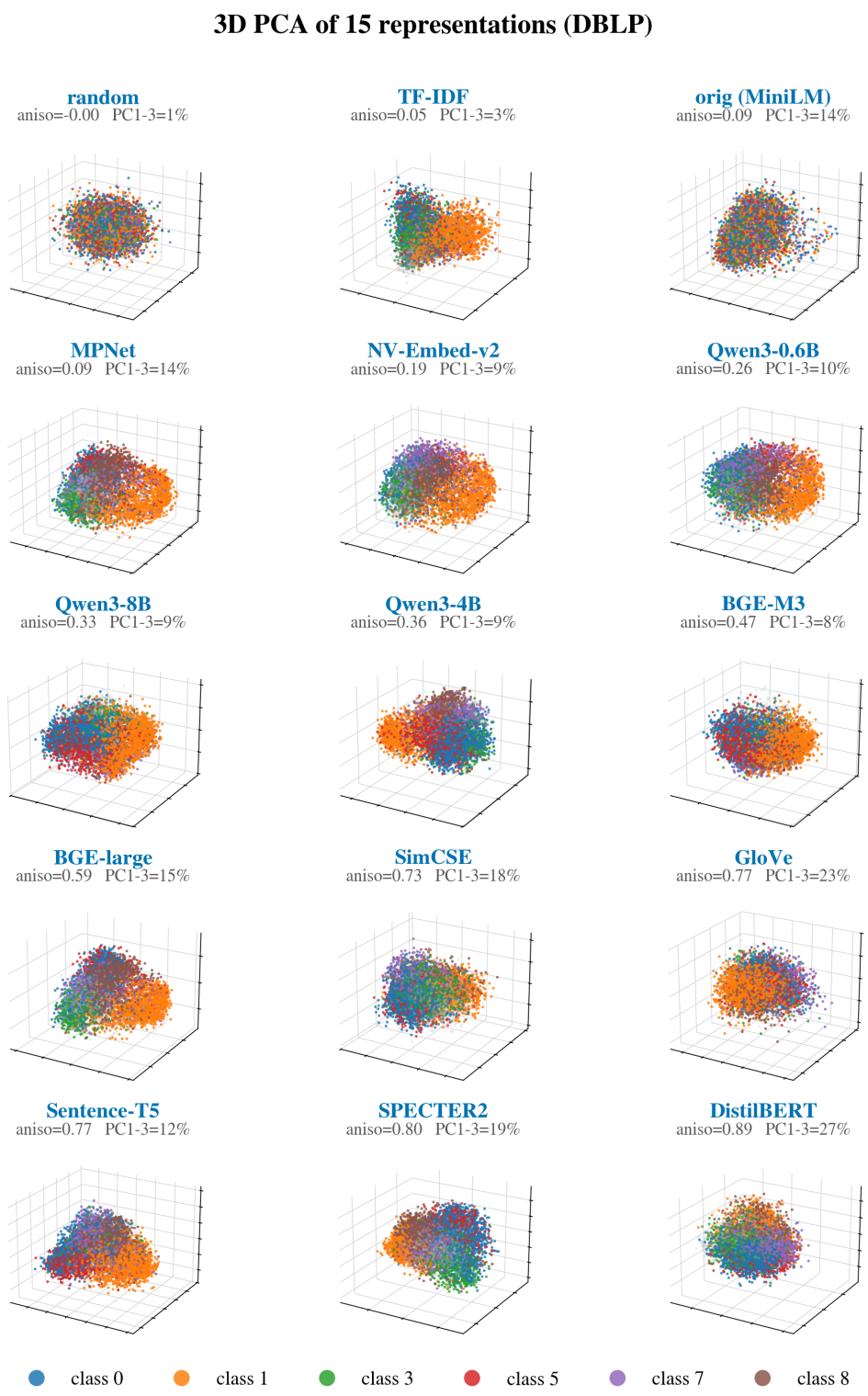


Figure 11: First three principal components of all 15 representations on DBLP. Layout, sampling, and annotations follow Figure 7.

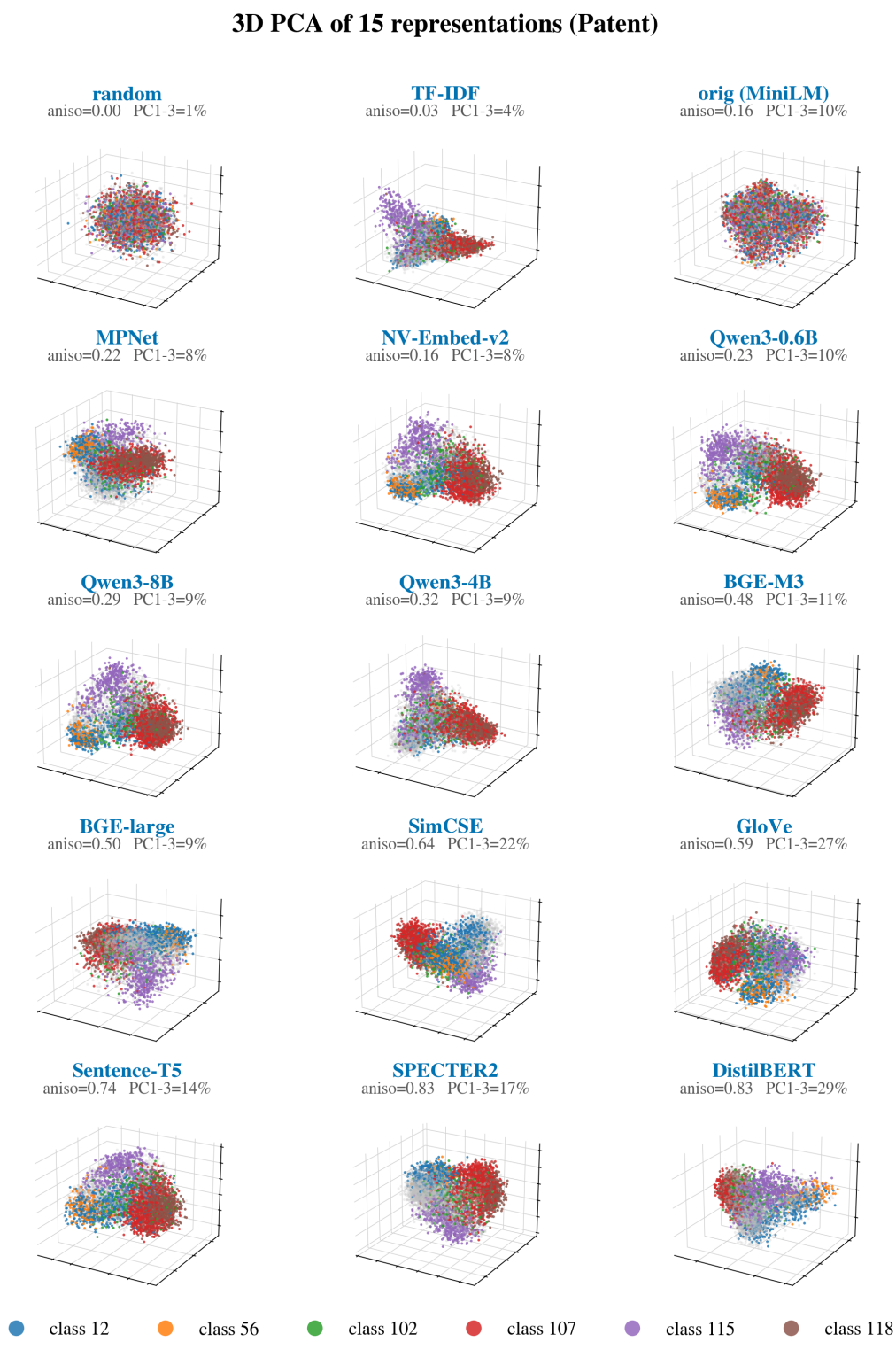


Figure 12: First three principal components of all 15 representations on Patent. Layout, sampling, and annotations follow Figure 7.

J LIMITATIONS AND FUTURE WORK

Our experiments establish a consistent association between TGA and residual homogeneous message-passing gains within HTAGs. In the evaluated settings, additional relation-aware gains are generally smaller and show less consistent associations with TGA, leaving their reliable prediction unresolved. TGA therefore supports prioritizing graph-model evaluation across candidate node representations, but does not provide calibrated estimates of gain magnitude or guarantee that message passing will improve performance. Future work should develop diagnostics sensitive to these smaller relation-aware gains and disentangle how raw text, graph topology, and representation quality jointly determine graph utility in HTAGs.