CoLIFT: Compatibility-Guided Graph-Free Distillation under Heterophily

Anonymous authors
Paper under double-blind review

Abstract

GNN-to-MLP distillation enables efficient graph-free inference while preserving rich information learned by GNNs. However, existing methods are developed primarily for homophilous graphs and rarely distinguish which neighborhood relations provide transferable supervision. This limitation becomes critical under heterophily, where observed neighborhoods mix structurally compatible relations with weakly supported or potentially misleading ones. To address this challenge, we propose CoLIFT, a **CO**mpatibility-guided **LIFT**ing framework that uses class compatibility as a structural proxy for GNN-to-MLP relation-level transferability. CoLIFT estimates compatibility from reliability-weighted teacher class-pair co-occurrences and corrects it for class prevalence. It then lifts these scores to node-pair sampling probabilities through capped exponential calibration that preserves each target node's expected relation budget. The sampled relations jointly define probability-weighted relation distillation and contrastive supervision that admits structurally supported cross-class positives. Across seven benchmarks, CoLIFT achieves the highest accuracy among the graph-free baselines, with a 6.00-percentage-point macro-average gain over the feature-only MLP on heterophilous graphs. Fixed-budget diagnostics and post-hoc analysis find that our calibration generally shifts probability mass toward structurally supported class pairs, with high-compatibility contexts yielding higher mean accuracy than low-compatibility ones on all seven datasets.

1 Introduction

Graph neural networks (GNNs) achieve strong performance by combining node attributes with graph topology through message passing (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018). However, their reliance on neighborhood aggregation complicates deployment. Predicting a node may require repeated access to its multi-hop neighborhood, which introduces computation, memory, and communication overhead. Knowledge distillation from GNNs to multilayer perceptrons (MLPs) addresses this limitation by transferring structural knowledge from a GNN teacher to an MLP student during training. Only the MLP is retained for graph-free inference (Zhang et al., 2022). Recent methods have enriched this paradigm through positional features, representation alignment, frequency decomposition, knowledge reliability, and structure-aware distillation targets (Tian et al., 2023; Wu et al., 2023a;b; Yang et al., 2024). Nevertheless, their designs and evaluations have largely centered on homophilous graphs, where neighboring nodes tend to provide mutually consistent supervision. This gap matters in settings such as word-dependency and product co-purchasing networks, where connections can link nodes with different labels (Platonov et al., 2023). It remains unclear how structural knowledge should be transferred when observed edges encode a mixture of relations with different semantic and structural roles.

Heterophily exposes this gap. Heterophily-aware architectures can better capture complex graph structures under such conditions (Zhu et al., 2020; Chien et al., 2021; Lim et al., 2021). These advances improve how GNNs exploit heterophilous structure, but they do not resolve how that structure should be transferred to an MLP student. Under a controlled setting with the teacher, student, data splits, and training protocol fixed, we find that representative graph-free distillation methods yield less stable gains on heterophilous graphs and may even underperform a feature-only MLP. This suggests a mismatch between structural utility for the teacher and transferability to the student. We refer to the underlying problem as relation-level transferability, which asks whether a relation exploited

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.

by the teacher can provide useful supervision to a graph-free student. Heterophilous neighborhoods may mix same-class relations, structurally compatible cross-class relations, and relations with weak structural support. Treating them uniformly can dilute useful knowledge and induce negative transfer.

To address this problem, we propose CoLIFT, a **CO**mpatibility-guided **LIFT**ing framework for graph-free distillation under heterophily. CoLIFT first estimates class-level structural compatibility from reliability-weighted teacher predictions and observed edges. It then lifts this evidence to node-pair transfer scores and softly calibrates the relation pairs and weights used for relation-level knowledge distillation while preserving each target node's expected relation budget. Finally, relation-aware contrastive distillation provides additional representation-level supervision. This mechanism changes the content of the distillation context without simply increasing its size. It also avoids brittle hard filtering and helps the student internalize compatible structural patterns. After training, all graph-dependent components are discarded, and only the MLP student is retained for inference.

Our main contributions are summarized as follows:

- We identify relation-level transferability as an overlooked problem in graph-free distillation under heterophily. Matched fixed-budget interventions isolate the effect of relation composition and show that compatibility provides a useful signal for transfer decisions.
- We introduce CoLIFT, whose compatibility-lifting mechanism estimates reliability-weighted, prior-corrected class compatibility and applies capped exponential calibration to a base relation proposal, yielding node-pair sampling probabilities while preserving each target's expected relation budget. The calibrated probabilities govern probability-weighted relation distillation and relation-aware contrastive supervision with structurally supported cross-class positives.
- Across seven benchmarks, CoLIFT improves over the feature-only MLP on every dataset and attains the highest mean accuracy among the evaluated shared-protocol graph-free methods. Ablations and controlled diagnostics further assess its components and relation-selection behavior.

2 Related Work

Graph-free GNN distillation. The Graph-less Neural Networks (GLNN) framework distills GNN predictions into an MLP for graph-free inference (Zhang et al., 2022). Subsequent work transfers positional and representation cues, frequency-specific knowledge, reliability estimates, sample hardness, or discrete structural codes, including Full-Frequency GNN-to-MLP (FF-G2M) distillation (Tian et al., 2023; Wu et al., 2023a;b; 2024b; Yang et al., 2024). AdaGMLP boosts student MLPs and aligns nodes under limited labels or incomplete features (Lu et al., 2024); SimMLP uses self-supervised alignment to encode local structure (Wang et al., 2025); and TAG2M develops task-agnostic structural transfer (V et al., 2025). Structure-Aware MLP (SA-MLP) incorporates structural embeddings, while Prototype-Guided Knowledge Distillation (PGKD) transfers class-prototype knowledge (Chen et al., 2024; Wu et al., 2025). TINED uses teacher injection and Dirichlet energy distillation, whereas Dual-Stream Aligned MLP (DA-MLP) aligns complementary graph signals (Zhou et al., 2025; Yu et al., 2026). Which teacher-exploited relations should be transferred under heterophily remains less clear.

Graph learning under heterophily. Cross-class connections are not uniformly harmful (Ma et al., 2022; Platonov et al., 2023). Prior work adapts neighborhoods, propagation, frequency responses, or feature–structure fusion (Abu-El-Haija et al., 2019; Pei et al., 2020; Zhu et al., 2020; Bo et al., 2021; Chien et al., 2021; Lim et al., 2021; Li et al., 2022; Luan et al., 2022). CPGNN uses class compatibility to guide graph propagation (Zhu et al., 2021). Edge-Splitting MLP (ES-MLP) separates edges for learning on homophilous and heterophilous graphs (Kohn et al., 2025), while HarmonyGNNs adapts structural modeling across both regimes (Xue & Wu, 2026). These approaches model diverse relations but do not directly address their transfer from a GNN teacher to an MLP student.

Relation-aware knowledge transfer. Relational distillation transfers pairwise geometry or contrastive dependencies between teacher and student representations (Park et al., 2019; Tung & Mori,

2019; Tian et al., 2020). Graph-MLP constructs contrastive supervision from adjacent nodes; LinkDist uses contrastive links; and Teacher-Free Graph Self-Distillation (TGS) uses neighborhood self-distillation (Hu et al., 2021; Luo et al., 2021; Wu et al., 2024a). These methods demonstrate the value of relational supervision but generally do not use class-pair compatibility to calibrate which relations enter teacher-to-student transfer. CoLIFT uses estimated compatibility to recalibrate node-level distillation contexts under a fixed expected relation budget and supervise the student at prediction and representation levels.

3 PRELIMINARIES

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with $N = |\mathcal{V}|$ nodes, adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$, and node features $\mathbf{X} \in \mathbb{R}^{N\times d}$. Each node belongs to one of C classes, while labels $y_i \in \{1,\dots,C\}$ are observed only for nodes in $\mathcal{V}_L$. In GNN-to-MLP distillation, a pretrained GNN teacher f_T accesses both features and graph structure to produce logits $\mathbf{Z}^T = f_T(\mathbf{X}, \mathbf{A})$, whereas a graph-free MLP student f_S produces $\mathbf{Z}^S = f_S(\mathbf{X})$. The student is optimized with $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}}$, where $\mathcal{L}_{\text{sup}}$ is the supervised classification loss over $\mathcal{V}_L$, $\mathcal{L}_{\text{KD}}$ matches the teacher's predictive distribution, and λ_{KD} controls the distillation strength. After training, only f_S is retained for graph-free inference.

We describe the graph-level label structure using edge homophily $h_{\text{edge}} = \frac{1}{|\mathcal{E}|}\sum_{(i,j)\in\mathcal{E}} \mathbb{I}(y_i = y_j)$, where $\mathbb{I}(\cdot)$ is the indicator function. A smaller h_{edge} indicates a larger proportion of cross-class edges, but this graph-level statistic does not reveal which class pairs are overrepresented relative to their background prevalence. We use it only to describe datasets and estimate class-pair compatibility separately in Section 5 as structural evidence for relation transfer.

4 MOTIVATION

Controlled transferability gap. A strong GNN teacher does not necessarily yield a strong graph-free student. We compare a feature-only MLP with a GLNN student distilled from a GPR-GNN teacher under fixed data splits, student architecture, and training protocol. On Texas, the teacher outperforms the MLP by 7.02 percentage points, yet GLNN remains comparable to the MLP; under the same setting on Cora, GLNN gains 25.10 points. Teacher quality alone therefore cannot explain distillation success. Although existing methods transfer predictions, representations, or structural cues, how the composition of teacher-exploited relations affects cross-architecture supervision remains insufficiently understood.

Relation-level transferability. Given a fixed teacher, student architecture, training protocol, and relation budget, we define relation-level transferability as the expected contribution of an observed relation or relation composition when used as distillation supervision for a graph-free student. This issue is particularly important under heterophily, where neighborhoods mix same-class relations, structurally supported cross-class relations, and weakly supported ones. Because transferability is an outcome property unavailable before training, we use class compatibility, which measures class-pair support relative to background prevalence, as an estimable structural proxy. Under matched relation budgets, changing only the relation composition alters student outcomes, while high-compatibility contexts yield higher mean accuracy than low-compatibility ones on all seven datasets (Section 6.4). Nevertheless, paired outcomes vary, indicating that class compatibility does not fully determine the benefit of every individual relation. These observations motivate a mechanism that preserves relation quantity, uses compatibility as soft evidence for node-level transfer decisions, and distills the resulting relations at both prediction and representation levels. CoLIFT instantiates this design.

5 PROPOSED MODEL

Figure 1 presents the three components of CoLIFT: reliability-aware compatibility discovery, budget-preserving compatibility lifting, and relation-aware contrastive distillation.

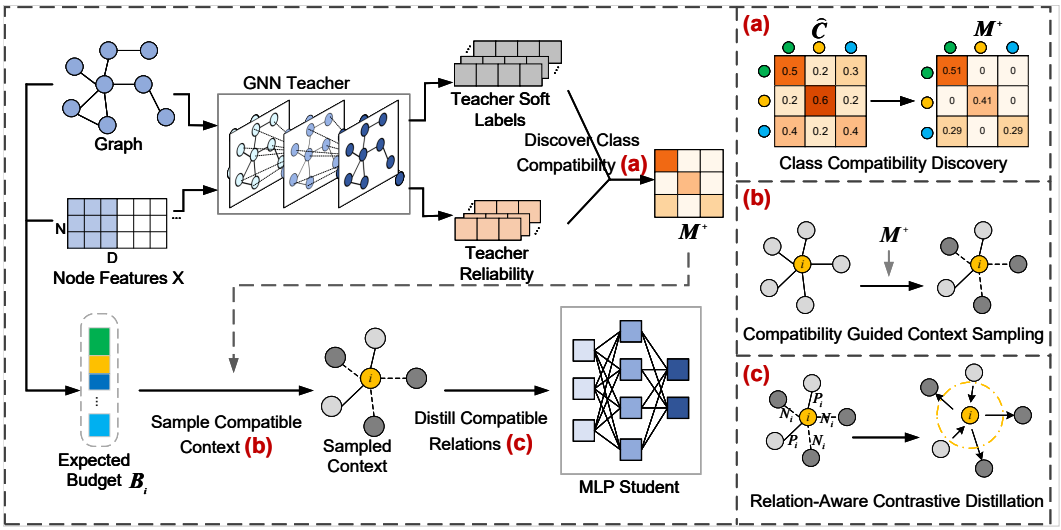

Figure 1: Overview of CoLIFT. Reliability-weighted teacher predictions and observed edges provide class-level compatibility evidence. CoLIFT lifts this evidence to node pairs, redistributes relation probabilities within each target node's expected budget, and organizes relation-aware contrastive supervision.

Reliability-Aware Compatibility Discovery. Building on prior class-compatibility modeling for graph propagation (Zhu et al., 2021), CoLIFT estimates class-pair support from teacher predictions and observed edges. Let $\mathbf{q}_i = \mathrm{softmax}(\mathbf{z}_i^T/T)$ be the temperature-scaled prediction of a pretrained and frozen GNN teacher. Since uncertain predictions can distort class-relation statistics, we weight each node by its normalized predictive confidence:

$$r_i = \mathrm{clip}\left(1 - \frac{H(\mathbf{q}_i)}{\log C}, r_{\min}, 1\right), \qquad H(\mathbf{q}_i) = -\sum_{c=1}^{C} q_{ic} \log q_{ic}. \tag{1}$$

We further anchor training nodes with their observed labels. Specifically, $\tilde{\mathbf{q}}_i = \mathrm{onehot}(y_i)$ for $v_i \in \mathcal{V}_L$ and $\tilde{\mathbf{q}}_i = \mathbf{q}_i$ otherwise. The resulting class prior is $\boldsymbol{\pi} = N^{-1}\sum_i \tilde{\mathbf{q}}_i$. Only training labels are used in this construction. We then aggregate reliability-weighted class co-occurrences along observed edges:

$$\widehat{\mathbf{C}} = \mathrm{RowNorm}\left(\sum_{(u,v)\in\mathcal{E}} r_u r_v \tilde{\mathbf{q}}_u \tilde{\mathbf{q}}_v^\top + \beta\mathbf{I} + \alpha\mathbf{1}\boldsymbol{\pi}^\top\right). \tag{2}$$

Here, $\widehat{C}_{ab}$ estimates the conditional preference for a node of class a to connect to class b. The terms $\alpha\mathbf{1}\boldsymbol{\pi}^\top$ and $\beta\mathbf{I}$ add prior-proportional and same-class pseudo-counts, respectively, to stabilize sparse estimates.

To remove the influence of class prevalence, we compare each conditional connection with the corresponding neighbor-class prior:

$$M^+_{ab} = \left[\log \frac{\widehat{C}_{ab} + \epsilon}{\pi_b + \epsilon}\right]_+. \tag{3}$$

A positive M^+_{ab} indicates that class pair (a, b) occurs more frequently than expected under the background class distribution. We term this reliability-weighted and prior-corrected class-pair association class compatibility. It serves as class-level structural evidence for transfer decisions and is subsequently lifted to node-level relations during student training.

Budget-Preserving Compatibility Lifting. The class-level matrix $\mathbf{M}^+$ does not directly determine which node-level relations should be transferred. For each target node v_i, let $\mathcal{C}_i$ contain its

one-hop, non-self-loop candidate neighbors. The base proposal $p_{ij}^{(0)}$ follows HGMD's hardness-aware neighbor probabilities (Wu et al., 2024b), with expected budget $\overline{B}_i = \sum_{j \in \mathcal{C}_i} p_{ij}^{(0)}$. CoLIFT preserves this budget while recalibrating context composition.

Let $\widehat{y}_i = \arg\max_c q_{ic}$ denote the teacher-predicted class. We lift class compatibility to each candidate relation through $s_{ij} = M^+_{\widehat{y}_i \widehat{y}_j}$ and recalibrate the base proposal using capped exponential calibration:

$$\begin{aligned} \ell_{ij} &= \log\Big(\max\{p_{ij}^{(0)}, \epsilon\}\Big) + \gamma \, \mathrm{Norm}_i(s_{ij}), \\ p_{ij} &= \min\{1, \kappa_i \exp(\ell_{ij})\}, \qquad \sum_{j \in \mathcal{C}_i} p_{ij} = \overline{B}_i. \end{aligned} \tag{4}$$

Here, $\mathrm{Norm}_i(\cdot)$ normalizes compatibility within the candidate set, and γ controls its influence. The scaling factor κ_i is chosen to satisfy the budget constraint. Allocations above one are capped, and the remaining mass is redistributed among uncapped candidates. This procedure guarantees valid Bernoulli probabilities while preserving the expected relation budget. It retains the base proposal but shifts probability toward structurally supported relations, allowing compatibility to act as soft evidence rather than a hard ranking rule.

Each candidate is then sampled independently as $m_{ij} \sim \mathrm{Bernoulli}(p_{ij})$, producing the local context $\mathcal{S}_i = \{j \mid m_{ij} = 1\}$. Individual contexts may vary across draws, but their expected size remains unchanged. The sampled relations and their calibrated probabilities directly define the pairs and weights used for relation-level knowledge distillation. Let $\widetilde{\mathcal{E}}$ contain the sampled relations and self-loops, and let c_{ij} denote the normalized calibrated weight for relation (i, j). With q_j^T and q_i^S denoting the temperature-scaled teacher and student distributions, respectively, we define

$$\mathcal{L}_{\mathrm{rKD}} = \frac{1}{|\widetilde{\mathcal{E}}|} \sum_{(i,j) \in \widetilde{\mathcal{E}}} c_{ij} \, \mathrm{KL}\big(q_j^T \parallel q_i^S\big). \tag{5}$$

Relation-Aware Contrastive Distillation. Relation-level knowledge distillation transfers teacher predictions along the calibrated relations, but does not explicitly organize the student representation. CoLIFT therefore uses the same compatibility evidence to provide relation-aware contrastive supervision. For each target node, prediction-consistent and structurally supported cross-class relations form the positive set, while the remaining sampled relations form the negative set:

$$\begin{aligned} \mathcal{P}_i &= \Big\{ j \in \mathcal{S}_i \;\Big|\; \widehat{y}_i = \widehat{y}_j \text{ or } M^+_{\widehat{y}_i \widehat{y}_j} > 0 \Big\}, \\ \mathcal{N}_i &= \mathcal{S}_i \setminus \mathcal{P}_i. \end{aligned} \tag{6}$$

Unlike supervised contrastive learning, which forms positives from shared labels (Khosla et al., 2020), this definition also admits cross-class relations with stable structural support. Low compatibility indicates limited positive structural support rather than an inherently harmful relation.

Let $\mathbf{h}_i$ denote the student representation and $\mathbf{g}_i = \mathrm{Proj}(\mathbf{h}_i)$ its training-only projection. Following the InfoNCE principle (van den Oord et al., 2018), CoLIFT applies a multi-positive contrastive loss in this projection space:

$$\mathcal{L}_{\mathrm{ctr},i} = -\log \frac{\displaystyle\sum_{j \in \mathcal{P}_i} \exp(\cos(\mathbf{g}_i, \mathbf{g}_j)/\tau)}{\displaystyle\sum_{j \in \mathcal{P}_i \cup \mathcal{N}_i} \exp(\cos(\mathbf{g}_i, \mathbf{g}_j)/\tau)}. \tag{7}$$

The loss brings the student closer to prediction-consistent and compatible cross-class relations relative to low-compatibility ones. The projection head limits direct interference with the classification space. We average the loss over target nodes with non-empty positive and negative sets.

The complete training objective is $\mathcal{L} = \lambda \mathcal{L}_{\mathrm{sup}} + (1 - \lambda)\mathcal{L}_{\mathrm{rKD}} + \lambda_{\mathrm{ctr}} \mathcal{L}_{\mathrm{ctr}}$, where $\lambda \in [0, 1]$ balances supervised learning and relation-level knowledge distillation. After training, all graph-dependent components and the projection head are removed. Only the MLP student is retained for graph-free inference.

6 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following questions. (1) Can CoLIFT improve graph-free node classification on both homophilous and heterophilous graphs? (2) How do its core components contribute to the overall performance? (3) Does relation composition affect transfer outcomes, and does CoLIFT shift its contexts toward relations with stronger structural support? (4) Can CoLIFT achieve fully graph-free inference with MLP-level efficiency?

6.1 EXPERIMENTAL SETUP

Datasets and splits. We evaluate CoLIFT on seven homophilous and heterophilous node-classification benchmarks. The homophilous citation networks are Cora, Citeseer, and PubMed (Sen et al., 2008). The heterophilous benchmarks are Texas, Wisconsin, and Actor (Pei et al., 2020), and Roman-empire (Platonov et al., 2023). Edge homophily ranges from 0.05 for Roman-empire to 0.81 for Cora, covering different class-mixing patterns. Dataset descriptions and statistics appear in Appendix C.1. Except for Roman-empire, we use one fixed split and ten seeds per dataset; Roman-empire uses ten official splits with one seed each. All controlled methods share the same split–seed pairs. Unless otherwise specified, we report mean test accuracy and standard deviation over ten runs.

Comparison protocol. We use GPR-GNN as the primary teacher and adopt the same MLP architecture for all controlled graph-free students. For each run, the teacher checkpoint is selected by validation accuracy and frozen during student training. Controlled distillation methods share its checkpoint and outputs within the matched run. They also follow identical preprocessing, data splits, random seeds, student architectures, and checkpoint-selection rules. Common hyperparameters use the same settings or candidate ranges, while method-specific choices are tuned on validation data only. Methods evaluated with their official implementations and independent protocols are marked as reference comparisons. Graph-dependent models are included only as performance references and are excluded from the ranking of graph-free methods.

Training and evaluation. We follow the transductive semi-supervised node-classification setting. During training, a student may access the complete node-feature matrix, the training graph, and soft predictions from the frozen teacher. Only training-node labels contribute to the supervised objective and compatibility estimation. Validation labels are used solely for hyperparameter tuning, early stopping, and checkpoint selection. Test labels remain inaccessible until the model configuration and checkpoint have been fixed. At inference, the trained MLP receives node features. The graph, teacher, compatibility estimator, relation sampler, and contrastive projection head are all discarded.

6.2 MAIN RESULTS

Table 1: Node classification accuracy (%, mean ± std.); group averages are macro-averages. Boldface and underlining mark the best and second-best shared-protocol graph-free results, respectively; † marks an unranked official-protocol result.

Method	Heterophilous Graphs					Homophilous Graphs			
	Texas	Wisconsin	Actor	Roman-empire	Avg. Het.	Cora	Citeseer	PubMed	Avg. Hom.
Feature-only Baseline									
MLP	75.68±3.42	74.12±4.00	34.89±0.61	66.01±0.58	62.68	59.26±1.68	57.56±1.96	72.30±1.30	63.04
Graph-based Models									
SAGE	51.35±8.20	42.35±2.35	30.05±0.50	52.36±0.47	44.03	82.62±0.72	68.88±0.92	76.86±0.74	76.12
GPR-GNN	82.70±3.67	72.16±6.00	35.96±0.39	71.25±0.40	65.52	83.34±0.69	70.92±0.55	79.48±0.22	77.91
Graph-free Distillation Baselines									
GLNN (SAGE)	62.70±4.95	65.09±3.37	31.67±0.71	61.81±0.48	55.32	83.18±0.39	69.78±0.53	79.26±0.48	77.41
HGMD (GCN)	62.70±5.77	67.65±5.63	34.76±0.82	58.23±0.86	55.84	83.72±1.15	68.48±1.51	80.48±0.54	77.56
NOSMOG†	60.81±7.56	60.98±10.56	29.26±2.16	46.79±1.55	49.46	83.04±1.26	73.78±1.54	77.34±2.36	78.05
Controlled GPR-GNN Distillation									
GLNN (GPR)	75.68±6.16	69.80±1.57	<u>36.75±0.49</u>	<u>67.25±0.36</u>	62.37	<u>84.36±0.71</u>	61.02±14.94	80.12±0.32	75.17
HGMD (GPR)	<u>85.89±2.48</u>	<u>82.16±2.83</u>	35.53±0.63	66.66±0.48	<u>67.56</u>	83.45±1.40	69.20±2.06	80.30±0.41	77.65
FF-G2M (GPR)	70.00±4.09	67.84±3.30	34.19±1.68	58.15±2.36	57.55	83.66±1.44	<u>71.21±0.85</u>	<u>81.46±0.36</u>	<u>78.78</u>
KRD (GPR)	70.81±4.15	71.76±4.49	36.26±0.65	66.56±0.48	61.35	83.76±1.65	70.24±1.29	81.11±0.34	78.37
PGKD (GPR)	77.03±5.59	76.27±5.43	35.51±1.33	66.48±0.49	63.82	83.62±1.30	71.14±0.88	80.74±0.33	78.50
CoLIFT (Ours)	**87.22±2.69**	**82.55±2.69**	**37.09±1.15**	**67.86±0.38**	**68.68**	**84.38±0.49**	**71.90±2.51**	**81.76±0.21**	**79.35**

Table 1 reports the node classification results across seven benchmarks. Among the graph-free methods evaluated under the shared protocol, CoLIFT achieves the highest accuracy on every dataset. It also consistently improves over the feature-only MLP, with no negative transfer observed after distillation. The macro-average gains over MLP reach 6.00 percentage points on heterophilous graphs and 16.31 points on homophilous graphs. These results show that CoLIFT can transfer graph-structural knowledge to a feature-only student across distinct relational regimes.

The heterophilous results are particularly relevant to our setting. With the GPR-GNN teacher, student architecture, data splits, random seeds, and model-selection protocol controlled, CoLIFT outperforms every controlled GPR-GNN distillation baseline on all four heterophilous datasets. On homophilous graphs, CoLIFT also uses same-class relation supervision and attains the highest shared-protocol macro-average, although its margins over the strongest baselines are modest (0.02–0.69 pp). With H2GCN as the teacher, CoLIFT also improves over GLNN on all seven datasets, with a 1.46 pp overall macro-average gain (Appendix C.2). CoLIFT further exceeds the reported NOSMOG accuracy on six of the seven datasets, although this comparison remains descriptive because NOSMOG follows an independent official protocol. Together, these results establish the effectiveness of compatibility-aware transfer without compromising graph-free inference. Sections 6.3 and 6.4 next examine the component contributions and relation-level mechanisms underlying these gains.

6.3 How Do Different Components Contribute to CoLIFT?

CoLIFT combines compatibility lifting (CL) with relation-aware contrastive distillation (RCD). We isolate their effects with *w/o CL* (RCD on uncalibrated contexts), *w/o RCD* (CL without the contrastive objective), and *w/o CCP* (same-class positives only on heterophilous graphs). All other settings remain unchanged.

Table 2: Component ablations in node classification accuracy (%, mean $\pm$ std.). CL, RCD, and CCP denote compatibility lifting, relation-aware contrastive distillation, and compatible cross-class positives. Each Δ is CoLIFT minus the corresponding ablation (pp); a dash means CCP is not used in the selected full configuration.

Dataset	w/o CL	w/o RCD	w/o CCP	CoLIFT	Δ_{CL}	Δ_{RCD}	Δ_{CCP}
Texas	87.30 ± 3.13	87.09 ± 2.48	87.57 ± 3.17	87.22 ± 2.69	-0.08	$+0.13$	-0.35
Wisconsin	82.16 ± 2.69	82.16 ± 2.55	81.76 ± 2.45	82.55 ± 2.69	$+0.39$	$+0.39$	$+0.79$
Actor	36.22 ± 1.08	36.29 ± 0.82	35.93 ± 0.67	37.09 ± 1.15	$+0.87$	$+0.80$	$+1.16$
Roman-empire	66.84 ± 0.36	66.87 ± 0.39	66.76 ± 0.42	67.86 ± 0.38	$+1.02$	$+0.99$	$+1.10$
Cora	83.64 ± 1.77	83.84 ± 1.34	–	84.38 ± 0.49	$+0.74$	$+0.54$	–
Citeseer	69.90 ± 2.09	69.71 ± 2.37	–	71.90 ± 2.51	$+2.00$	$+2.19$	–
PubMed	80.98 ± 0.51	80.99 ± 0.48	–	81.76 ± 0.21	$+0.78$	$+0.77$	–

Table 2 shows the component-wise ablation results on all seven datasets. Removing CL reduces accuracy on six datasets, while removing RCD degrades performance on every dataset. The two components therefore play complementary roles. CL reshapes the relational context presented to the student, whereas RCD converts that context into representation-level supervision. Compatible cross-class positives provide further gains on three of the four heterophilous datasets, suggesting that transferable relations need not be restricted to same-class pairs. Texas is the exception: full CoLIFT is 0.08 pp below w/o CL and 0.35 pp below w/o CCP, indicating that compatibility evidence does not guarantee a mean gain in every setting. These findings support the layered design of CoLIFT and motivate the mechanism analysis in Section 6.4.

6.4 How Does Compatibility Lifting Reshape Distillation Contexts?

To isolate relation composition from relation quantity, we compare *Random*, *Homophily*, *Compatibility-High*, and *Compatibility-Low* under the same per-node budget of $k = 2$. Paired runs share the data split, frozen teacher outputs, student initialization, and candidate sets; only the selected relations differ. This restricted setting is a mechanism diagnostic rather than an ablation or performance reproduction of full CoLIFT.

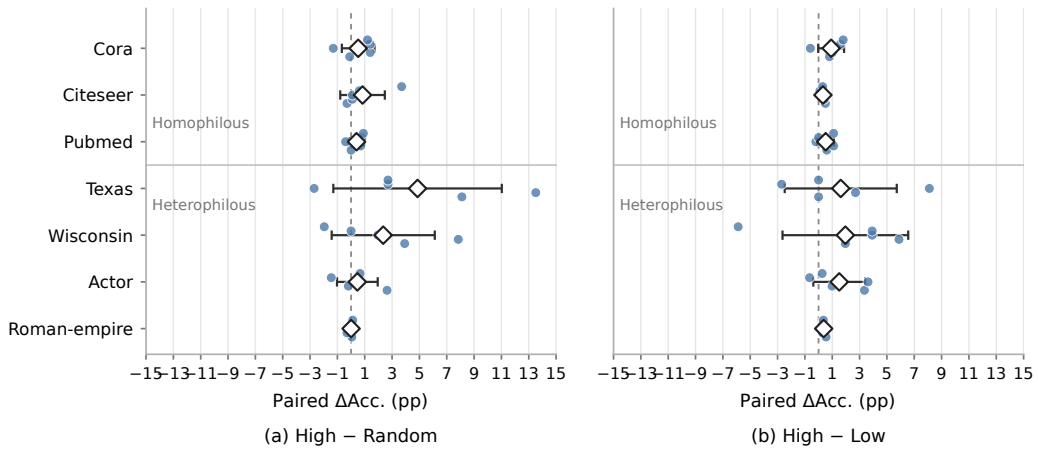

Figure 2: Paired test-accuracy differences under fixed-budget relation interventions with $k = 2$. Panels (a) and (b) report Compatibility-High minus Random and Compatibility-High minus Compatibility-Low, respectively. All other conditions are shared within each paired run. Filled circles denote five paired runs, while open diamonds and error bars show the mean and sample standard deviation. ΔAcc. is reported in percentage points (pp).

Figure 2 shows that Compatibility-High achieves higher mean accuracy than both Random and Compatibility-Low on all seven datasets, although its advantage over Random is negligible on Roman-empire. Across heterophilous graphs, the paired gains average 1.920 and 1.368 pp, respectively, versus 0.587 pp for both contrasts on homophilous graphs. Wisconsin nevertheless has both positive and negative paired differences. These controlled results show that relation composition affects transfer under a fixed budget and that compatibility provides an informative, but variable, ranking signal.

For full CoLIFT, we aggregate base and calibrated sampling probabilities over training rounds for validation and test target nodes. An oracle compatibility matrix M^{oracle} is constructed after training from complete labels and the observed graph. For sampling probabilities p, we define the probability-weighted oracle relation quality as $Q_{\text{oracle}}(p) = \frac{\sum_e \sum_{(i,j)} p_{ij}^{(e)} M^{\text{oracle}}_{y_i y_j}}{\sum_e \sum_{(i,j)} p_{ij}^{(e)}}$, where the sums cover all recorded rounds and evaluated candidate relations. We also measure changes in the probability mass assigned to oracle positive-compatible relations and their cross-class subset. The oracle quantities are post-hoc and do not affect probability calibration, relation sampling, model optimization, or checkpoint selection. The candidate sets and expected relation budgets remain unchanged.

Table 3: Relation quality before and after CoLIFT's soft probability calibration. Pre/post values are means over five runs; changes are post minus pre (mean ± sample std.). Q_{oracle} denotes probability-weighted oracle compatibility. ΔM^{+} and $\Delta M^{+}_{\text{cross}}$ denote changes in probability mass on oracle positive-compatible relations and their cross-class subset (pp). Interpret Q_{oracle} within each dataset.

Dataset	Q_{oracle}: Pre → Post	ΔQ_{oracle}	ΔM_{+} (pp)	$\Delta M^{+}_{\text{cross}}$ (pp)
Cora	$1.354 \to 1.365$	0.011 ± 0.003	0.644 ± 0.181	0.000 ± 0.000
Citeseer	$1.053 \to 1.055$	0.002 ± 0.001	0.151 ± 0.066	-0.001 ± 0.008
PubMed	$0.652 \to 0.654$	0.002 ± 0.000	0.200 ± 0.024	0.000 ± 0.000
Texas	$0.235 \to 0.255$	0.019 ± 0.016	-0.002 ± 1.613	0.683 ± 1.849
Wisconsin	$0.246 \to 0.248$	0.001 ± 0.003	0.922 ± 2.437	0.200 ± 2.090
Actor	$0.021 \to 0.021$	≈ 0.000	0.157 ± 0.111	0.010 ± 0.164
Roman-empire	$0.387 \to 0.430$	0.043 ± 0.002	1.206 ± 0.142	1.755 ± 0.093

Table 3 shows positive mean ΔQ_{oracle} on six datasets, led by Roman-empire (0.043 ± 0.002), where compatible cross-class mass increases by 1.755 ± 0.093 pp. Wisconsin's change is small and variable (0.001 ± 0.003), while Actor's is negligible.

The fixed-budget intervention links relation composition to accuracy, while the probability analysis shows how full CoLIFT reallocates its context budget. These diagnostics differ from the ablations in Table 2: Q_{oracle} measures class-pair structural support, whereas accuracy reflects the full student-training objective. Actor's positive ablation differences despite negligible ΔQ_{oracle} illustrate this distinction. Compatibility is therefore a structural proxy rather than a sufficient measure of transferability; low compatibility indicates limited positive support, not an inherently harmful relation.

6.5 Can CoLIFT Preserve Efficient Graph-Free Inference?

CoLIFT exploits graph structure during training, but deploys only the MLP student. We therefore examine whether its structural transfer gains introduce additional inference costs. We report predictive performance and deployment efficiency on Roman-empire, the largest graph in our benchmark.

Table 4: Predictive performance and inference efficiency on Roman-empire. Bold and underlined values denote the best and second-best graph-free methods, respectively.

Method	Accuracy (%)	# Params.	Time (ms)	Memory (MiB)	Graph
GPR-GNN	$71.25_{\pm 0.40}$	20,445	$8.975_{\pm 0.020}$	69.50	Yes
MLP	$66.01_{\pm 0.58}$	426,002	$1.542_{\pm 0.021}$	170.27	No
GLNN (GPR)	$\underline{67.25_{\pm 0.36}}$	426,002	$1.541_{\pm 0.016}$	170.27	No
HGMD(GPR)	$66.66_{\pm 0.48}$	426,002	$1.542_{\pm 0.013}$	170.27	No
CoLIFT	$\mathbf{67.86_{\pm 0.38}}$	426,002	$1.565_{\pm 0.010}$	170.27	No

Accuracy is averaged over the ten official splits. Inference time is measured on the same GPU after 50 warm-up runs and reported over five blocks of 200 full-node inference runs. Parameter count and peak GPU memory include only the deployed model and the inputs required at inference.

As shown in Table 4, CoLIFT achieves the highest accuracy among all graph-free methods. It improves over the same-capacity MLP by 1.85 percentage points and over the strongest competing distillation baseline, GLNN (GPR), by 0.61 points. These gains require no additional deployment parameters or peak GPU memory, and the inference latency remains comparable to that of the other MLP students. After training, CoLIFT discards the teacher, graph structure, compatibility matrix, sampled relations, and projection head. Inference therefore requires only a feature-based MLP forward pass. Although the narrower GPR-GNN uses fewer parameters and less memory, it still performs graph propagation at inference. CoLIFT is consequently $5.73\times$ faster while remaining entirely graph-free. These results show that CoLIFT internalizes relation-level knowledge without increasing the deployment cost of the shared student backbone.

7 Conclusion

In this paper, we study relation-level transferability in graph-free distillation under heterophily and propose CoLIFT. CoLIFT introduces a compatibility-lifting mechanism that estimates reliability-weighted, prior-corrected class compatibility and recalibrates a base relation proposal into node-pair sampling probabilities via capped exponential calibration while preserving each target node's expected relation budget. The sampled relations define relation-level knowledge distillation and relation-aware contrastive supervision, including structurally supported cross-class positives. On seven benchmarks, CoLIFT improves over feature-only MLPs and attains the highest mean accuracy among the evaluated shared-protocol graph-free methods. Fixed-budget interventions show that relation composition can affect student accuracy, while post-hoc analysis characterizes changes in class-pair structural support under calibration. After training, CoLIFT discards all graph-dependent components and retains efficient MLP-only inference. These findings support compatibility-guided graph-free distillation in the studied settings.

8 AI USE STATEMENT

In this work, we used OpenAI ChatGPT (GPT-5.6 Sol) solely for language editing to improve the grammar, clarity, and readability of the manuscript. We did not use generative AI for research ideation, methodology or experiment design, method implementation, data analysis, result interpretation, mathematical reasoning, or literature review. We reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 21–29. PMLR, 2019. URL https://proceedings.mlr.press/v97/abu-el-haija19a.html.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 3950–3957, 2021. doi: 10.1609/aaai.v35i5.16514.

Jie Chen, Mingyuan Bai, Shouzhen Chen, Junbin Gao, Junping Zhang, and Jian Pu. SA-MLP: Distilling graph knowledge from GNNs into structure-aware MLP. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=MZ2kKZc8m7.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations*, 2021.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1024–1034, 2017.

Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. Graph-MLP: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*, 2021. URL https://arxiv.org/abs/2106.04051.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

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

Matthias Kohn, Marcel Hoffmann, and Ansgar Scherp. Edge-splitting MLP: Node classification on homophilic and heterophilic graphs without message passing. In *Proceedings of the Third Learning on Graphs Conference*, volume 269 of *Proceedings of Machine Learning Research*, pp. 11:1–11:21. PMLR, 2025. URL https://proceedings.mlr.press/v269/kohn25a.html.

Xiang Li, Renyu Zhu, Yao Cheng, Caihua Shan, Siqiang Luo, Dongsheng Li, and Weining Qian. Finding global homophily in graph neural networks when meeting heterophily. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13242–13256. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22ad.html.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20887–20902, 2021.

Weigang Lu, Ziyu Guan, Wei Zhao, and Yaming Yang. AdaGMLP: AdaBoosting GNN-to-MLP knowledge distillation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2060–2071, 2024. doi: 10.1145/3637528.3671699.

Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1362–1375, 2022.

Yi Luo, Aiguo Chen, Ke Yan, and Ling Tian. Distilling self-knowledge from contrastive links to classify graph nodes without passing messages. *arXiv preprint arXiv:2106.08541*, 2021. URL https://arxiv.org/abs/2106.08541.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? In *International Conference on Learning Representations*, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations*, 2023.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: 10.1609/aimag.v29i3.2157.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V. Chawla. Learning MLPs on graphs: A unified view of effectiveness, robustness, and efficiency. In *International Conference on Learning Representations*, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

Ram Ganesh V, Ayush Singh, Aditi Rai, Harsh Pal, Deepanshu Bagotia, Akshay Sethi, Aakarsh Malhotra, and Sayan Ranu. TAG2M: A task-agnostic knowledge distillation framework for distilling GNNs to MLPs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2859–2870, 2025. doi: 10.1145/3711896.3737145.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

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

Zehong Wang, Zheyuan Zhang, Chuxu Zhang, and Yanfang Ye. Training MLPs on graphs without supervision. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 697–706, 2025. doi: 10.1145/3701551.3703550.

Lirong Wu, Haitao Lin, Yufei Huang, Tianyu Fan, and Stan Z. Li. Extracting low-/high-frequency knowledge from graph neural networks and injecting it into MLPs: An effective GNN-to-MLP distillation framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9): 10351–10360, 2023a. doi: 10.1609/aaai.v37i9.26232.

Lirong Wu, Haitao Lin, Yufei Huang, and Stan Z. Li. Quantifying the knowledge in GNNs for reliable distillation into MLPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37571–37581. PMLR, 2023b. URL https://proceedings.mlr.press/v202/wu23m.html.

Lirong Wu, Haitao Lin, Zhangyang Gao, Guojiang Zhao, and Stan Z. Li. A teacher-free graph knowledge distillation framework with dual self-distillation. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4375–4385, 2024a. doi: 10.1109/TKDE.2024.3374773.

Lirong Wu, Yunfan Liu, Haitao Lin, Yufei Huang, and Stan Z. Li. Teach harder, learn poorer: Rethinking hard sample distillation for GNN-to-MLP knowledge distillation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2554–2563, 2024b. doi: 10.1145/3627673.3679586.

Taiqiang Wu, Zhe Zhao, Jiahao Wang, Xingyu Bai, Lei Wang, Ngai Wong, and Yujiu Yang. Edge-free but structure-aware: Prototype-guided knowledge distillation from GNNs to MLPs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5656–5667. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.379/.

Rui Xue and Tianfu Wu. HarmonyGNNs: Harmonizing heterophily and homophily in GNNs via self-supervised node encoding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c33176723974e9b50ca0297d769903c8-Abstract-Conference.html.

Ling Yang, Ye Tian, Minkai Xu, Zhongyi Liu, Shenda Hong, Wei Qu, Wentao Zhang, Bin Cui, Muhan Zhang, and Jure Leskovec. VQGraph: Rethinking graph representation space for bridging GNNs and MLPs. In *International Conference on Learning Representations*, 2024.

Zhiyuan Yu, Mingkai Lin, Wenzhong Li, Zhangyue Yin, Shijian Xiao, and Sanglu Lu. Demystifying GNN-to-MLP knowledge transfer: Theoretical grounding and dual-stream distillation method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33):28005–28013, 2026. doi: 10.1609/aaai.v40i33.40025.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old MLPs new tricks via distillation. In *International Conference on Learning Representations*, 2022.

Ziang Zhou, Zhihao Ding, Jieming Shi, Qing Li, and Shiqi Shen. TINED: GNNs-to-MLPs by teacher injection and Dirichlet energy distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78616–78632. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25a.html.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7793–7804, 2020.

Jiong Zhu, Ryan A. Rossi, Anup Rao, Tung Mai, Nedim Lipka, Nesreen K. Ahmed, and Danai Koutra. Graph neural networks with heterophily. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11168–11176, 2021. doi: 10.1609/aaai.v35i12.17332.

A Pseudocode of Compatibility-Guided Relation Transfer

Algorithm 1 summarizes the training procedure. Class compatibility is estimated once from the frozen teacher and the observed graph. Relation probabilities are then recalibrated during student training. Only the selected MLP is used at inference.

Algorithm 1 Training CoLIFT

Require: Graph $G = (\mathcal{V}, \mathcal{E})$, features $\mathbf{X}$, training labels $\{y_i : i \in \mathcal{V}_L\}$, frozen teacher f_T, student MLP f_S, and training hyperparameters with $\alpha, \beta \geq 0$ and $\alpha + \beta > 0$
Ensure: Validation-selected, graph-free student $f_S(\cdot; \theta^\star)$

Compute $\mathbf{q}_i = \mathrm{softmax}(\mathbf{z}_i^T/T)$, $\widehat{y}_i = \arg\max_c q_{ic}$, and $r_i = \mathrm{clip}(1 - H(\mathbf{q}_i)/\log C, r_{\min}, 1)$ for all i.
Set $\widetilde{\mathbf{q}}_i = \mathrm{onehot}(y_i)$ if $i \in \mathcal{V}_L$; otherwise set $\widetilde{\mathbf{q}}_i = \mathbf{q}_i$.
Compute $\boldsymbol{\pi} = |\mathcal{V}|^{-1} \sum_{i \in \mathcal{V}} \widetilde{\mathbf{q}}_i$ and $\mathbf{W} = \sum_{(u,v) \in \mathcal{E}} r_u r_v \widetilde{\mathbf{q}}_u \widetilde{\mathbf{q}}_v^\top$.
Set $\widehat{\mathbf{C}} = \mathrm{RowNorm}(\mathbf{W} + \beta \mathbf{I} + \alpha \mathbf{1}\boldsymbol{\pi}^\top)$.
Set $M^+_{ab} = \left[\log \frac{\widehat{C}_{ab} + \epsilon}{\pi_b + \epsilon}\right]_+$ for all class pairs (a, b).
for each student-training epoch **do**
  Compute current student predictions and representations.
  for each target node $i \in \mathcal{V}$ **do**
    Let $\mathcal{C}_i$ be the one-hop, non-self-loop candidates.
    if $\mathcal{C}_i = \varnothing$ **then**
      Set $\mathcal{S}_i = \mathcal{P}_i = \mathcal{N}_i = \varnothing$.
      continue
    end if
    Obtain base probabilities $p_{ij}^{(0)}$ for $j \in \mathcal{C}_i$; set $\overline{B}_i = \sum_{j \in \mathcal{C}_i} p_{ij}^{(0)}$.
    Compute $s_{ij} = M^+_{\widehat{y}_i \widehat{y}_j}$ and $\ell_{ij} = \log \max\{p_{ij}^{(0)}, \epsilon\} + \gamma\, \mathrm{Norm}_i(s_{ij})$ for $j \in \mathcal{C}_i$.
    if $\overline{B}_i \in \{0, |\mathcal{C}_i|\}$ **then**
      Set $p_{ij} = \overline{B}_i / |\mathcal{C}_i|$ for every $j \in \mathcal{C}_i$.
    else
      Find $\kappa_i > 0$ such that $\sum_{j \in \mathcal{C}_i} \min\{1, \kappa_i e^{\ell_{ij}}\} = \overline{B}_i$.
      Set $p_{ij} = \min\{1, \kappa_i e^{\ell_{ij}}\}$ for every $j \in \mathcal{C}_i$.
    end if
    Independently draw $m_{ij} \sim \mathrm{Bernoulli}(p_{ij})$; set $\mathcal{S}_i = \{j \in \mathcal{C}_i : m_{ij} = 1\}$.
    Set $\mathcal{P}_i = \{j \in \mathcal{S}_i : \widehat{y}_i = \widehat{y}_j \text{ or } M^+_{\widehat{y}_i \widehat{y}_j} > 0\}$.
    Set $\mathcal{N}_i = \mathcal{S}_i \setminus \mathcal{P}_i$.
  end for
  Add one deterministic self-loop per target outside $\mathcal{C}_i$ and $\overline{B}_i$; form normalized KD weights c_{ij} over the sampled relations and self-loops.
  Compute $\mathcal{L}_{\mathrm{sup}}$ on training labels, $\mathcal{L}_{\mathrm{rKD}}$ on sampled relations and self-loops, and $\mathcal{L}_{\mathrm{ctr}}$ from valid $(\mathcal{P}_i, \mathcal{N}_i)$.
  Update the student and projection head using $\mathcal{L} = \lambda \mathcal{L}_{\mathrm{sup}} + (1 - \lambda)\mathcal{L}_{\mathrm{rKD}} + \lambda_{\mathrm{ctr}} \mathcal{L}_{\mathrm{ctr}}$.
  Save the student checkpoint with the best validation accuracy.
end for
return the saved MLP $f_S(\cdot; \theta^\star)$.

B Theoretical Analysis

We analyze the statistical meaning of class compatibility and the properties of budget-preserving calibration. Prior correction and smoothing determine how class-pair support is measured. Capped calibration then reallocates relation probabilities under a fixed expected budget. We show that this allocation has a generalized KL projection interpretation and responds monotonically to compatibility strength. Finally, independent relation sampling preserves the expected context size. All results concern the operators defined in Section 5.

B.1 Statistical Interpretation of Class Compatibility

Assume $C \geq 2$, $\alpha, \beta \geq 0$, and $\epsilon > 0$. Using the same edge orientation convention as Eq. (2), define the reliability-weighted soft counts

$$W_{ab} = \sum_{(u,v)\in\mathcal{E}} r_u r_v \widetilde{q}_{ua} \widetilde{q}_{vb}, \qquad W_a = \sum_{b=1}^{C} W_{ab}. \tag{8}$$

Here, $\widetilde{\mathbf{q}}_i$ includes training-label anchoring. The reliability r_i is still computed from the original teacher prediction $\mathbf{q}_i$ through Eq. (1). Since $\sum_b \pi_b = 1$, Eq. (2) becomes

$$\widehat{C}_{ab} = \frac{W_{ab} + \beta \mathbb{I}(a = b) + \alpha \pi_b}{W_a + \beta + \alpha}. \tag{9}$$

We consider rows with $W_a + \beta + \alpha > 0$.

Proposition B.1 (Prior correction and smoothing). The deviation from the node-class prior satisfies

$$\widehat{C}_{ab} - \pi_b = \frac{W_{ab} - W_a \pi_b + \beta\big(\mathbb{I}(a = b) - \pi_b\big)}{W_a + \beta + \alpha}. \tag{10}$$

Thus, $M^+_{ab} > 0$ if and only if the numerator in Eq. (10) is positive. For fixed W, $\boldsymbol{\pi}$, and β, increasing α preserves the sign of this deviation and reduces or maintains its magnitude. Positive compatibility scores therefore shrink toward zero. For cross-class relations, $a \neq b$,

$$M^+_{ab} > 0 \quad \Longleftrightarrow \quad W_{ab} > (W_a + \beta)\pi_b. \tag{11}$$

Proof. Subtracting π_b from Eqs. (9) and (10) gives the stated deviation. Eq. (3) adds the same positive ϵ to its numerator and denominator. Its logarithm is therefore positive exactly when $\widehat{C}_{ab} > \pi_b$. The denominator in Eq. (10) is positive, so the numerator determines the sign. Increasing α enlarges only this denominator. The sign is unchanged, and the absolute deviation cannot increase. For a positive deviation, $\widehat{C}_{ab}$ approaches π_b, so monotonicity of the logarithm gives the stated shrinkage. Setting $a \neq b$ yields Eq. (11). □

The two smoothing terms have distinct effects. The prior term $\alpha \mathbf{1}\boldsymbol{\pi}^\top$ shrinks conditional statistics toward background prevalence. With other statistics fixed, it cannot create positive compatibility on its own. The diagonal term βI favors same-class relations. It also raises the threshold for positive cross-class compatibility. A positive cross-class score thus requires sufficient reliability-weighted association to exceed this threshold.

An idealized case gives a direct probability interpretation. On a fixed graph, suppose the original teacher predictions satisfy $\mathbf{q}_i = \mathrm{onehot}(y_i)$ for every node. With $0 \leq r_{\min} \leq 1$, Eq. (1) then gives $r_i = 1$. Set $\alpha = \beta = 0$. Let P_E denote the endpoint-class distribution of a uniformly sampled ordered edge, and let P_V denote the class distribution of a uniformly sampled node. For $W_a > 0$ and $\pi_b > 0$,

$$\lim_{\epsilon \downarrow 0} M^+_{ab} = \left[\log \frac{P_E(Y_v = b \mid Y_u = a)}{P_V(Y = b)}\right]_+. \tag{12}$$

Indeed, W_{ab}/W_a is exactly the empirical conditional endpoint frequency, while π_b is the empirical node-class proportion. When the conditional frequency is zero, the right-hand side is defined as zero by continuity of the clipped expression. This identity requires no independence assumption on graph edges. It does not assert that soft teacher statistics consistently estimate the true class distribution. The reference distribution is node prevalence; edge-endpoint marginals may differ because of degree heterogeneity.

These results explain the structural evidence represented by M^+. Beyond distinguishing same-class and cross-class relations, it measures class-pair support relative to background prevalence. CoLIFT uses this evidence as a structural proxy for relation-level transferability. The next section characterizes how it changes relation probabilities. Teacher errors, local variation, and the student objective still affect transfer outcomes. Structural support alone therefore does not guarantee a gain from an individual relation. The fixed-budget interventions in Section 6.4 assess the proxy empirically at the level of controlled context-selection strategies.

B.2 Optimality of Capped Budget-Preserving Calibration

For a target node i, fix the candidate set $\mathcal{C}_i$, base probabilities $p_{ij}^{(0)} \in [0,1]$, and finite normalized scores $t_{ij} = \mathrm{Norm}_i(s_{ij})$. Define

$$\begin{aligned} n_i &= |\mathcal{C}_i|, & \overline{B}_i &= \sum_{j \in \mathcal{C}_i} p_{ij}^{(0)}, \\ a_{ij} &= \max\{p_{ij}^{(0)}, \epsilon\} > 0, & w_{ij}(\gamma) &= a_{ij} e^{\gamma t_{ij}}. \end{aligned} \tag{13}$$

These quantities are fixed for one calibration operation. The base probabilities need not remain constant throughout student training. Consider the generalized KL projection

$$\begin{aligned} \min_{\mathbf{p}_i} \quad & D(\mathbf{p}_i \| \mathbf{w}_i) = \sum_{j \in \mathcal{C}_i} \left[p_{ij} \log \frac{p_{ij}}{w_{ij}} - p_{ij} + w_{ij} \right] \\ \text{subject to} \quad & 0 \le p_{ij} \le 1, \qquad \sum_{j \in \mathcal{C}_i} p_{ij} = \overline{B}_i, \end{aligned} \tag{14}$$

with the convention $0 \log(0/w) = 0$.

Proposition B.2 (Generalized KL projection). For $n_i > 0$, $0 \le \overline{B}_i \le n_i$, and finite γ, Eq. (14) has a unique minimizer $\mathbf{p}_i^\star(\gamma)$. If $0 < \overline{B}_i < n_i$, there is a unique $\kappa_i > 0$ such that

$$p_{ij}^\star(\gamma) = \min\{1, \kappa_i w_{ij}(\gamma)\}, \qquad \sum_{j \in \mathcal{C}_i} p_{ij}^\star(\gamma) = \overline{B}_i. \tag{15}$$

Proof. The feasible set is nonempty, compact, and convex. For example, setting each coordinate to $\overline{B}_i / n_i$ is feasible. The objective is continuous and strictly convex on the nonnegative orthant because every $w_{ij} > 0$. A unique minimizer therefore exists.

For $0 < \overline{B}_i < n_i$, no optimal coordinate is zero. Otherwise, moving a sufficiently small mass from a positive coordinate to a zero coordinate would decrease the objective; this follows from the right derivative tending to $-\infty$ at zero. The lower bounds are thus inactive at the optimum. Let ν be the budget multiplier and $\mu_{ij} \ge 0$ the upper-bound multipliers. The KKT conditions give

$$\log \frac{p_{ij}}{w_{ij}} + \nu + \mu_{ij} = 0, \qquad \mu_{ij}(p_{ij} - 1) = 0. \tag{16}$$

An uncapped coordinate has $\mu_{ij} = 0$ and hence $p_{ij} = e^{-\nu} w_{ij}$. A capped coordinate satisfies $p_{ij} = 1$ and $\mu_{ij} = \log(e^{-\nu} w_{ij}) \ge 0$. Setting $\kappa_i = e^{-\nu}$ gives Eq. (15). The function $F_i(\kappa) = \sum_j \min\{1, \kappa w_{ij}\}$ is continuous, with $F_i(0) = 0$. It is strictly increasing until it reaches n_i. Every interior budget therefore determines a unique positive κ_i. □

When $\overline{B}_i = 0$, the solution is the all-zero vector and $\kappa_i = 0$ is valid. When $\overline{B}_i = n_i$, the solution is the all-one vector. Any $\kappa_i \ge \max_j w_{ij}^{-1}$ then satisfies the budget, so the multiplier is not unique. An empty candidate set requires no calibration. In all cases, the optimal probability vector is unique. For an interior budget and finite scores, every candidate retains a positive probability. Compatibility thus guides soft reallocation within the candidate set rather than assigning exact zeros through hard filtering.

Proposition B.2 also gives a regularization interpretation. Let $\mathcal{F}_i$ denote the feasible set in Eq. (14). Dropping terms independent of $\mathbf{p}_i$ yields the equivalent problem

$$\min_{\mathbf{p}_i \in \mathcal{F}_i} D(\mathbf{p}_i \| \mathbf{a}_i) - \gamma \sum_j p_{ij} t_{ij}. \tag{17}$$

For $\gamma \ge 0$, calibration balances deviation from the base weights against a preference for higher compatibility scores. The separate budget constraint controls the expected relation count. This optimality concerns probability allocation, not student risk.

Corollary B.1 (Monotone compatibility preference). Fix the candidate set, base probabilities, budget, and normalized scores. For $\overline{B}_i > 0$, define

$$Q_i^{\mathrm{score}}(\gamma) = \frac{1}{\overline{B}_i} \sum_j p_{ij}^\star(\gamma) t_{ij}. \tag{18}$$

For any finite $\gamma_2 > \gamma_1$,

$$Q_i^{\mathrm{score}}(\gamma_2) \geq Q_i^{\mathrm{score}}(\gamma_1). \tag{19}$$

If $p_{ij}^{(0)} \geq \epsilon$ for every candidate, then $\mathbf{p}_i^\star(0) = \mathbf{p}_i^{(0)}$. For any $\gamma \geq 0$, the calibrated probability-weighted mean score is therefore no lower than that of the base proposal.

Proof. Write $D_k = D(\mathbf{p}_i^\star(\gamma_k) \| \mathbf{a}_i)$ and $T_k = \sum_j p_{ij}^\star(\gamma_k) t_{ij}$. Optimality at each parameter value gives

$$D_2 - \gamma_2 T_2 \leq D_1 - \gamma_2 T_1, \qquad D_1 - \gamma_1 T_1 \leq D_2 - \gamma_1 T_2.$$

Adding these inequalities yields $(\gamma_2 - \gamma_1)(T_2 - T_1) \geq 0$. Since $\gamma_2 > \gamma_1$ and $\overline{B}_i > 0$, Eq. (19) follows. If all base probabilities are at least ϵ, then $\mathbf{a}_i = \mathbf{p}_i^{(0)}$ is feasible. At $\gamma = 0$, it uniquely attains zero generalized KL divergence. Thus $\mathbf{p}_i^\star(0) = \mathbf{p}_i^{(0)}$. □

The corollary holds even when some probabilities reach the upper bound. It directly characterizes compatibility-guided budget reallocation. For two uncapped candidates, Eq. (15) also gives $p_{ij}^\star / p_{ik}^\star = (a_{ij}/a_{ik}) \exp\{\gamma(t_{ij} - t_{ik})\}$. If normalization has the fixed positive affine form $t_{ij} = A_i s_{ij} + d_i$, with $A_i > 0$, the same monotonicity holds for the probability-weighted mean of the original scores s_{ij}. Without this condition, the stated guarantee concerns t_{ij}.

If some base probabilities fall below ϵ, then $\mathbf{p}_i^\star(0)$ is the budget-constrained projection of the floored weights. It need not equal the original proposal. The monotonic comparison then starts from $\mathbf{p}_i^\star(0)$. Moreover, Q_i^{score} uses the normalized scores that drive calibration. It differs from the post-hoc oracle quantity Q_{oracle} in Section 6.4. The corollary establishes how calibration responds to its input signal. Whether oracle structural support and student performance improve remains an empirical question.

B.3 Expected Budget Preservation

For one sampling round, fix the candidate set and calibrated probabilities. Draw $m_{ij} \sim \mathrm{Bernoulli}(p_{ij}^\star)$ independently and define $\mathcal{S}_i = \{j \in \mathcal{C}_i : m_{ij} = 1\}$. All expectations and variances below are conditional on these fixed quantities.

Proposition B.3 (Expected budget and variance). For $n_i > 0$,

$$\begin{aligned} \mathbb{E}[|\mathcal{S}_i|] &= \overline{B}_i, \\ \mathrm{Var}(|\mathcal{S}_i|) &= \sum_j p_{ij}^\star (1 - p_{ij}^\star) \leq \overline{B}_i \left(1 - \frac{\overline{B}_i}{n_i}\right). \end{aligned} \tag{20}$$

Proof. Linearity of expectation and the budget constraint give $\mathbb{E}[|\mathcal{S}_i|] = \sum_j p_{ij}^\star = \overline{B}_i$. Independence gives $\mathrm{Var}(|\mathcal{S}_i|) = \overline{B}_i - \sum_j (p_{ij}^\star)^2$. The bound follows from Cauchy–Schwarz: $\sum_j (p_{ij}^\star)^2 \geq \overline{B}_i^2 / n_i$. □

Calibration therefore changes relation composition at a fixed expected size. The realized size can vary across rounds, with variance determined by the probability allocation. This result concerns non-self-loop candidates. The self-loops added in Eq. (5) are excluded from $\overline{B}_i$. If each target receives one fixed self-loop, its expected total context size is $\overline{B}_i + 1$, with unchanged variance. The static interventions in Section 6.4 fix the selected relation count. Full CoLIFT instead preserves the base proposal's expected relation budget through Bernoulli sampling.

C Additional Experimental Details and Results

C.1 Dataset Descriptions and Statistics

We evaluate CoLIFT on seven node-classification benchmarks covering both homophilous and heterophilous graphs.

- **Citation networks.** Cora, Citeseer, and PubMed are citation networks in which nodes represent scientific publications and edges represent citation relations. Node features are derived from textual content, while node labels indicate publication topics. Cora and Citeseer use sparse bag-of-words features. PubMed uses TF–IDF features and categorizes publications into three diabetes-related classes.
- **WebKB networks.** Texas and Wisconsin are university webpage networks from the WebKB collection. Nodes represent webpages, edges correspond to hyperlinks, and node features are bag-of-words representations. Each webpage belongs to one of five categories. Both datasets exhibit strong heterophily.
- **Actor network.** Actor is an actor co-occurrence network extracted from Wikipedia. Nodes represent actors, and edges connect actors that co-occur on the same webpage. Node features are derived from keywords in the corresponding Wikipedia pages. The nodes are divided into five classes.
- **Roman-empire network.** Roman-empire is a heterophilous graph from the Heterophilous Graph Benchmark. It contains 18 node classes and has the lowest edge homophily among the datasets considered in this work.

Table 5 reports the dataset statistics used in our experiments. Edge counts follow the preprocessing and counting conventions of our experimental pipeline.

Table 5: Dataset statistics. Lower edge homophily h_{edge} indicates stronger heterophily.

Dataset	Texas	Wisconsin	Actor	Roman-empire	Cora	Citeseer	PubMed
Classes	5	5	5	18	7	6	3
Features	1,703	1,703	932	300	1,433	3,703	500
Nodes	183	251	7,600	22,662	2,708	3,327	19,717
Edges	295	466	26,752	32,927	5,278	4,552	44,324
h_{edge}	0.11	0.21	0.22	0.05	0.81	0.74	0.80

C.2 Distillation with an H2GCN Teacher

To examine whether CoLIFT's gains are specific to the primary GPR-GNN teacher, we further evaluate it with H2GCN (Zhu et al., 2020), a GNN designed for learning under heterophily. For each run, an H2GCN teacher is selected by validation accuracy and then frozen during student training. Within each run, GLNN and CoLIFT share the same teacher checkpoint and outputs, preprocessing, data split, random seed, student architecture and initialization, and checkpoint-selection rules. Test labels remain inaccessible until the model configuration and checkpoint have been fixed. The Δ_{MLP} and Δ_{GLNN} columns report absolute differences between the mean accuracy of CoLIFT and that of the corresponding graph-free baseline in percentage points (pp).

Table 6: Test accuracy (%, mean ± std.) with H2GCN as the teacher. H2GCN is graph-dependent, while MLP, GLNN, and CoLIFT are graph-free at inference. Δ_{MLP} and Δ_{GLNN} denote differences between CoLIFT and the corresponding baselines, computed from reported mean accuracies in percentage points (pp). Group averages are macro-averages. Boldface marks the best graph-free result.

Dataset	MLP	H2GCN	GLNN (H2GCN)	CoLIFT (H2GCN)	Δ_{MLP}	Δ_{GLNN}
Texas	75.68±3.42	80.81±1.99	80.81±1.99	**82.03±1.73**	↑ 6.35 pp	↑ 1.22 pp
Wisconsin	**74.12±4.00**	67.45±3.09	64.90±3.97	69.61±3.95	↓ 4.51 pp	↑ 4.71 pp
Actor	34.89±0.61	35.50±0.69	35.95±0.71	**36.76±0.65**	↑ 1.87 pp	↑ 0.81 pp
Roman-empire	66.01±0.58	68.53±0.47	67.40±0.59	**67.76±0.48**	↑ 1.75 pp	↑ 0.36 pp
Cora	59.26±1.68	75.48±0.78	78.54±0.75	**79.08±0.49**	↑ 19.82 pp	↑ 0.54 pp
Citeseer	57.56±1.96	62.16±1.22	64.76±1.56	**66.39±1.10**	↑ 8.83 pp	↑ 1.63 pp
PubMed	72.30±1.30	77.30±0.71	82.08±0.73	**83.05±0.89**	↑ 10.75 pp	↑ 0.97 pp
Heterophilous Avg.	62.68	63.07	62.27	**64.04**	↑ 1.37 pp	↑ 1.78 pp
Homophilous Avg.	63.04	71.65	75.13	**76.17**	↑ 13.13 pp	↑ 1.05 pp
Overall Avg.	62.83	66.75	67.78	**69.24**	↑ 6.41 pp	↑ 1.46 pp

As shown in Table 6, CoLIFT achieves higher mean accuracy than GLNN with the same H2GCN teacher on all seven datasets. The improvements range from 0.36 pp on Roman-empire to 4.71 pp on Wisconsin. Averaged over the four heterophilous datasets, CoLIFT exceeds GLNN by 1.78 pp, compared with 1.05 pp on the three homophilous datasets and 1.46 pp overall. CoLIFT also outperforms the feature-only MLP on six datasets. Wisconsin is the exception, where H2GCN itself performs below the MLP under the evaluated protocol. These controlled results indicate that the benefit of compatibility-aware transfer is not specific to the primary GPR-GNN teacher, while its magnitude remains dataset-dependent.

C.3 Complete Fixed-Budget Relation Interventions

Tables 7 and 8 report the run-level test accuracies underlying the fixed-budget intervention in the main text. The four strategies share the same per-target relation budget of $k = 2$ within each paired run. Runs also share the data split, frozen teacher outputs, student initialization, and candidate sets. For Cora, Citeseer, PubMed, Texas, Wisconsin, and Actor, the run index equals the random seed. Roman-empire uses seeds 1000–1004 for run indices 0–4.

Table 7: Run-level test accuracy (%) for the fixed-budget intervention on homophilous datasets. High and Low denote Compatibility-High and Compatibility-Low, respectively.

Dataset	Run	Random	Homophily	High	Low
Cora	0	83.9	84.1	83.8	83.0
	1	82.8	84.2	84.2	83.2
	2	83.2	83.0	81.9	82.5
	3	82.7	83.5	84.1	82.5
	4	82.2	83.3	83.4	81.6
Citeseer	0	69.6	71.1	69.3	68.8
	1	70.9	71.3	71.0	70.6
	2	71.0	70.5	71.1	70.8
	3	71.5	70.6	72.1	72.0
	4	66.1	67.3	69.8	69.5
PubMed	0	81.1	81.2	81.1	80.5
	1	80.2	80.0	80.9	79.8
	2	81.2	80.9	80.8	81.0
	3	80.5	80.7	81.3	81.3
	4	80.2	81.6	81.1	80.0

Table 8: Run-level test accuracy (%) for the fixed-budget intervention on heterophilous datasets. High and Low denote Compatibility-High and Compatibility-Low, respectively.

Dataset	Run	Random	Homophily	High	Low
Texas	0	70.3	73.0	78.4	78.4
	1	62.2	73.0	75.7	73.0
	2	81.1	67.6	78.4	70.3
	3	73.0	73.0	75.7	78.4
	4	73.0	75.7	75.7	75.7
Wisconsin	0	70.6	74.5	74.5	72.5
	1	72.5	72.5	80.4	74.5
	2	74.5	68.6	76.5	72.5
	3	68.6	76.5	68.6	64.7
	4	64.7	64.7	62.7	68.6
Actor	0	34.3	35.6	36.9	33.6
	1	36.4	35.8	36.2	35.2
	2	35.9	36.7	36.5	32.9
	3	36.0	36.8	34.5	35.2
	4	34.9	35.2	35.6	35.3
Roman-empire	0	67.0	67.0	67.0	66.5
	1	67.5	67.1	67.2	66.9
	2	66.6	66.6	66.7	66.3
	3	66.1	65.2	66.2	65.9
	4	66.2	66.6	66.3	66.0

The run-level results clarify the variability behind the mean contrasts. Compatibility-High exceeds Random on average across all seven datasets, but its mean advantage is negligible on Roman-empire. Individual paired runs can also favor Random or Compatibility-Low. Moreover, Homophily has a slightly higher mean accuracy than Compatibility-High on Cora and Actor. The intervention therefore supports the effect of relation composition under a controlled budget, while showing that compatibility is an informative but imperfect ranking signal.