UNDERSTANDING THE ROLE OF LOCAL HOMOPHILY ON GNN PERFORMANCE

Kamila Abdiyeva, Brandon Baez, Javier Diego

ABSTRACT

Recent studies suggest that global graph properties can inform architecture selection and predict Graph Neural Network (GNN) performance. Our experiments, in line with prior work, reaffirm that *label homophily*, the tendency of connected nodes to share the same label, is one of the strongest predictors of GNN accuracy on node-classification tasks. Nevertheless, it remains unclear how message passing mechanistically exploits homophily to make individual predictions, and whether variation/misalignment in local node properties (compared to the global graph property) affects the resulting representations and recognition performance.

In this work, we systematically evaluate the role of homophily at both the *global* (graph-level) and *local* (node-level) scales, quantifying its effect on node representations and predictive performance. First, we compare the standard GNN to a label-based majority-vote baseline (NLM), isolating how much of the GNN performance can be attributed to simple local label agreement. As a contrasting extreme, we analyze a feature-only MLP, which removes the homophily effect entirely, allowing us to position GNN performance between these two baselines and assess the unique contribution of message passing. Next, we compare node embeddings throughout the message-passing process to study how homophily influences clustering in representation space and, in turn, node-level predictions. Finally, we perform controlled structural perturbations that locally decrease homophily and observe the resulting changes in classification accuracy.

1 INTRODUCTION

Graph Neural Networks (GNNs) have become the de facto standard for learning on relational data, largely due to their ability to propagate information across graph neighborhoods through message passing. A growing body of work has examined how *global graph properties* influence GNN performance, particularly in node-classification tasks. For instance, Procházka et al. (2023) showed that both node and edge homophily, the tendency of connected nodes to share the same label, are dominant drivers of GNN accuracy, far outweighing factors such as class imbalance or degree distribution. Building on this, more fine-grained analyses have explored the role of *local homophily*. Loveland et al. (2024) demonstrated that misclassifications frequently occur in neighborhoods where local homophily diverges from the global homophily of the graph, suggesting that global averages alone cannot explain node-level behavior. While these studies highlight the importance of homophilic structure, most remain *correlational*: they do not directly analyze the layer-wise predictions or representation dynamics of GNNs across different homophily regimes.

At the same time, it is well established that GNN performance often degrades under *heterophily*, where neighboring nodes have dissimilar labels (Zhu et al., 2020). However, more recent work (Ma et al., 2021) has shown that with appropriate architectural choices and hyperparameter tuning, GNNs can still achieve competitive performance even in heterophilous settings, indicating that homophily is not an absolute requirement for effective message passing.

In parallel, research in *GNN explainability* has introduced methods for identifying influential edges, subgraphs, and features behind a model's prediction. Techniques such as GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), and CF-GNNExplainer (Lucic et al., 2021) reveal the structural components that contribute to individual predictions. However, these approaches do not explain how message passing *systematically* transforms node representations under varying levels of homophily, nor whether message passing provides predictive benefits beyond simpler heuristics such as local label agreement.

Our work aims to bridge these two lines of research. We combine structural analysis, baseline comparison, representation dynamics, and targeted graph perturbations to study how *global and lo-*

cal homophily jointly shape GNN behavior. Rather than measuring correlations alone, we investigate the mechanistic role of homophily in message passing:

- When does message passing actually improve prediction over node features alone?
- When does local label agreement already explain GNN accuracy?
- How do embeddings move toward or away from homophilous neighbors across layers?
- How much homophily must be removed before a correct prediction flips?

Using these questions as motivation, we develop a unified framework connecting the *structural properties of graphs*, the behavior of *simple non-GNN baselines*, and the *internal embedding dynamics* of GNNs. This perspective provides a more mechanistic understanding of how and when message passing succeeds or fails, and clarifies the structural conditions under which GNNs offer benefits beyond node features or local label agreement. Our findings are as follows:

(1) Comparison between Global and Local Graph Statistics. Our error analysis reveals a consistent trend across all datasets: test nodes correctly classified by GNNs inhabit neighborhoods that are slightly more homophilous than the global graph mean, while nodes that GNNs fail to classify reside in neighborhoods that are significantly more heterophilous.

(2) Majority-vote baselines outperform GNNs. Across a range of homophilic datasets, we observe that a simple label-based majority-vote classifier frequently matches or exceeds the accuracy of message-passing GNNs. This indicates that a portion of GNN performance on homophilic graphs can be explained by local label agreement alone, without requiring learned aggregation or nonlinear transformations.

(3) Message passing offers limited advantages over feature-only models. We observed that for most highly homophilous datasets, GNNs outperformed the MLP baseline (with the exception of two models where the MLP achieved performance comparable to the GNN). In contrast, on heterophilous datasets, the MLP matched or even outperformed the GNN-based models. Notably, a feature-only MLP baseline proves competitive with message-passing GNNs across a significant portion of the benchmarks.

(4) Message passing pulls embeddings toward homophilous neighbors. By tracking representations across layers, we find that correctly classified nodes undergo systematic embedding drift toward the embeddings of their homophilous neighbors. This effect is significantly weaker or absent for incorrectly classified nodes or nodes in heterophilic neighborhoods. These results demonstrate that message passing does not merely smooth features globally, but selectively aligns embeddings with locally supportive neighborhood structure.

(5) Causal perturbation experiments on local homophily show that GNN predictions remain stable only when homophilous support is high but collapse under medium-high relative homophily reductions in both homophilous and heterophilous graphs, revealing a fundamental sensitivity of message passing to neighborhood label alignment.

2 Related Work

2.1 Homophily and Message Passing

Graph Neural Networks (GNNs) rely on message passing to aggregate information from a node's neighbors, and early architectures such as GCN (Kipf & Welling, 2017) and GAT (Velickovic et al., 2018) formalized this process using spectral and attention-based operators. Subsequent work has highlighted several limitations of message passing, including oversmoothing (Li et al., 2018; Oono & Suzuki, 2020) and bottlenecks in neighborhood aggregation (Alon & Yahav, 2021). In this work, we take a deeper look into the *dynamics* of message passing itself, examining how embeddings move across layers and how structural perturbations affect predictions. By introducing a directional drift metric and controlled homophily-reduction experiments, we aim to uncover the mechanisms behind these bottlenecks and characterize when message passing helps, when it harms, and why.

A central theme in understanding GNN performance is the role of *homophily*. Extensive empirical studies show that label homophily correlates strongly with GNN accuracy (Pei et al., 2020; Zhu et al., 2020). More granular analyses have revealed that misclassifications often arise in neighborhoods whose local homophily deviates from the global structure (Loveland et al., 2024), suggesting that homophily shapes the effectiveness of message passing at both global and local scales. However, most of these works are correlational; they do not investigate how message passing transforms node

embeddings in homophilous versus heterophilous neighborhoods. Our work addresses this gap by examining the *mechanistic* interaction between homophily and message passing through embedding drift, class-level clustering, and controlled perturbations.

2.2 Heterophily and Structural Adaptation

While homophily is often treated as a prerequisite for GNN success, recent works challenge this assumption. Ma et al. (2021) and others show that GNNs can remain competitive under heterophily by leveraging higher-order neighborhoods or decoupling feature and structural channels. These findings motivate a deeper understanding of *why* message passing sometimes succeeds or fails under different homophily conditions.

A parallel line of research develops architectures explicitly designed for heterophilous graphs, such as Geom-GCN (Pei et al., 2020), H2GCN (Zhu et al., 2020), MixHop (Abu-El-Haija et al., 2019), and GPR-GNN (Chamberlain et al., 2021). Although these models mitigate performance decay under heterophily, they focus primarily on architectural innovation rather than analyzing the representation-level behavior of message passing. By contrast, our work examines the embedding trajectories induced by message passing itself, clarifying when such architectures provide genuine representational advantages.

2.3 Explainability and Representation Drift

Explainability methods such as GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), and CF-GNNExplainer (Lucic et al., 2021) shed light on prediction-level dependencies by identifying influential edges or substructures. However, these techniques do not characterize the *directional effect* of message passing on node embeddings.

Our analysis introduces a structured notion of *message-passing drift*, which quantifies how embeddings move toward homophilous or heterophilous neighborhoods across layers. This perspective extends explainability by focusing on *representation dynamics*, revealing whether correct or incorrect predictions correspond to alignment or misalignment with local homophily. Furthermore, our homophily-reduction perturbation experiments provide causal evidence of how structural changes impact classification.

Relation to Feature Smoothing. Simple Graph Convolution (SGC) (Wu et al., 2019) interprets message passing as low-pass filtering of node features. Our findings refine this view: drift patterns show that effective message passing is not merely smoothing but a directional pull toward homophilous neighbors and away from heterophilous ones. This directional perspective connects structural conditions to the actual movement of embeddings.

2.4 Disentangling Features from Structure

Recent work emphasizes the importance of non-GNN baselines. Studies show that feature-only MLPs or nearest-neighbor heuristics can rival or outperform GNNs in low-homophily settings (Platonov et al., 2023). These findings align with our comparisons of MLP, node-label majority vote baselines, and GNNs. Our results refine this line of inquiry by explicitly quantifying when performance arises from (i) node features, (ii) label agreement, or (iii) genuine message-passing benefits.

2.5 Advances in Homophily Metrics and Theoretical Perspectives

Beyond scalar homophily, recent theory investigates finer structural factors. Zheng et al. (2024) argue that existing metrics do not capture the interplay among topology, labels, and features, proposing "Tri-Hom" to unify these perspectives. Luan et al. (2023) introduce "Node Distinguishability" to formalize when GNNs can succeed even under heterophily, showing that structural conditions can compensate for low homophily if inter-class feature separation is sufficiently strong.

Methodologically, Lu et al. (2025) use Large Language Models to estimate homophily from semantic cues, integrating LLM-derived graph priors into spectral filtering. Zheng et al. (2022) further classify heterophilous GNN architectures based on mixing patterns and label-sensitive aggregation.

Positioning Our Contribution. While these advances expand the understanding of homophily, heterophily, and GNN design, our work bridges the gap between metric-based predictions and empirical mechanism. By connecting homophily levels to (i) baseline comparisons, (ii) embedding drift, and (iii) perturbation outcomes, we provide a geometric and causal characterization of how message passing interacts with structural properties at both global and local scales.

3 Methodology

Given a graph $G = (V, E, X)$ with $n = |V|$ nodes, $m = |E|$ edges, and node features $X \in \mathbb{R}^{n \times d}$, we study how structural and label-dependent properties of the graph affect the behavior of message-

passing Graph Neural Networks (GNNs). Each node $v \in V$ has a label $y_v \in \mathcal{C}$ and feature vector x_v.

Our analysis proceeds through the following interconnected stages that jointly examine *global structure*, *local homophily*, *message passing dynamics*, and *causal perturbations*:

1. **Global structural characterization.** We first compute global graph statistics (Table 4 summarizes the metrics), such as label homophily, average node degree, clustering coefficient, and feature-level similarity, to quantify the large-scale structural properties of G and relate them to overall GNN performance.
2. **Baseline comparisons disentangling structural vs. feature signals.** We compare GNNs to two non-message-passing baselines:
   (a) a label-based majority-vote classifier (NLM) (as defined in Definition 7) that depends solely on neighborhood labels,
   (b) a feature-only multilayer perceptron (MLP) (as defined in Definition 8) that ignores graph structure.

   These comparisons isolate the predictive contributions of homophily and message passing.
3. **Representation dynamics under message passing.** We analyze how node embeddings evolve across GNN layers and investigate how local homophily affects the clustering, drift, and separability of learned representations. This reveals how neighborhood structure shapes GNN decisions.
4. **Causal perturbation of local homophily.** Finally, we introduce controlled modifications to local neighborhoods that increase or decrease homophily and measure their impact on node representations and predictions. These experiments provide causal evidence linking local homophily to GNN performance and failure modes.

Exact definition of the metrics discussed can be found in Appendix B.

4 Experiments

This section presents our empirical evaluation following the five-stage methodology described earlier. For clarity, we organize the experiments according to the global–local homophily framework and the role of message passing in shaping node representations.

4.1 Experimental Setup

Datasets. We evaluate our methods on thirteen widely used benchmark datasets covering different domains and homophily levels, described on Appendix C and summarized in Table 1. These datasets span a range of homophily levels, graph densities, and feature modalities, enabling us to evaluate how global and local homophily shape GNN performance and representation learning.

Table 1: Dataset statistics: number of nodes, edges, classes, feature dimensions, and global homophily.

Dataset	Nodes	Edges	Classes	Features	Homophily
Questions	48,921	153,540	10	300	0.8396
Cora	2,708	5,429	7	1,433	0.8100
Coauthor-CS	18,333	81,894	15	6,805	0.8081
PubMed	19,717	44,338	3	500	0.8024
Amazon-Computers	13,381	245,778	10	767	0.7772
CiteSeer	3,327	4,732	6	3,703	0.7355
Minesweeper	10,000	78,804	2	7	0.6828
Tolokers	11,758	1,038,000	2	10	0.5945
Amazon-ratings	24,492	186,100	5	300	0.3804
Chameleon	2,277	31,421	5	2,325	0.2350
Actor	7,600	30,019	5	932	0.2239
Squirrel	5,201	198,493	5	2,089	0.2188
Roman-empire	22,662	65,854	18	300	0.0469

Models. We evaluate five models covering both message-passing and non-message-passing paradigms. For GNNs, we use two-layer GCN (Kipf & Welling, 2017), GAT (Velickovic et al.,

2018), and GPRGNN (Chien et al., 2021) with ReLU activations as representative spectral, attention-based, and heterophily-robust architectures. To disentangle the role of structure, we compare against two baselines: (i) a Node-Label Majority (NLM) classifier, which predicts using the most common label in each node's 1-hop neighborhood, and (ii) a feature-only MLP, a two-layer fully connected network without access to graph edges. All models are trained with cross-entropy loss and identical early-stopping criteria, with shared hyperparameters (learning rate, hidden dimension, weight decay, dropout) held consistent across datasets.

4.2 Experiment 1.1: Global Structure vs. GNN Performance

Goal. Measure classification accuracy for Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Generalized PageRank Networks (GPRGNN) for node classification tasks defined on the datasets previously mentioned, and understand how global graph statistics correlate with accuracy on them.

Method. For each dataset, we train GCNs, GATs, and GPRGNNs with 64 hidden dimensions and for 200 epochs, and for 3 different seed values. We couldn't increase the number of seeds further due to computational constraints. We summarize the accuracy results in Table 2. Furthermore, we compute correlation coefficients between accuracy and the defined global statistics in Table 3.

Group	Dataset	Homophily η	GCN acc.	GAT acc.	GPRGNN acc.
Homophilous	Questions	0.8396	97.02 ± 0.00	97.02 ± 0.01	97.06 ± 0.00
	Cora	0.8100	81.03 ± 0.57	80.30 ± 1.15	79.80 ± 3.56
	Coauthor-CS	0.8081	94.84 ± 1.61	94.19 ± 1.40	96.50 ± 1.61
	PubMed	0.8024	78.83 ± 0.67	77.30 ± 0.17	79.23 ± 0.32
	Amazon-Computers	0.7772	90.19 ± 0.76	91.15 ± 0.91	90.70 ± 0.45
	CiteSeer	0.7355	68.53 ± 1.10	67.13 ± 1.55	70.17 ± 0.67
Heterophilous	Minesweeper	0.6828	80.49 ± 0.12	80.00 ± 0.00	82.67 ± 0.24
	Tolokers	0.5945	78.78 ± 0.24	78.06 ± 0.12	78.50 ± 0.06
	Amazon-ratings	0.3804	43.00 ± 0.54	41.52 ± 0.20	43.93 ± 0.15
	Chameleon	0.2350	38.23 ± 7.09	46.05 ± 0.96	42.25 ± 1.98
	Squirrel	0.2239	24.02 ± 2.18	29.43 ± 0.73	30.32 ± 1.53
	Actor	0.2188	28.09 ± 2.11	27.89 ± 0.39	34.45 ± 0.51
	Roman-empire	0.0469	45.50 ± 1.76	40.40 ± 0.33	68.80 ± 0.45

Table 2: Global homophily η and test accuracy (% $\pm$ s.d. over 3 seeds) of GCN, GAT, and GPRGNN on homophilous and heterophilous benchmarks.

Results. The accuracy results in Table 2 show a clear trend: GNN performance decreases as graph homophily η decreases (with few exceptions). This observation is reinforced by the correlation analysis in Table 3, where homophily exhibits the strongest Pearson correlation with test accuracy among all global statistics considered.

Model	η (homophily)	$\bar{d}$ (avg. degree)	C (global clustering)	s (attr. similarity)
GCN	0.917	0.038	-0.365	0.486
GAT	0.920	0.068	-0.341	0.425
GPRGNN	0.794	-0.008	-0.308	0.439

Table 3: Pearson correlation coefficients between global graph statistics and test accuracy for GCN, GAT, and GPRGNN across all datasets. It is calculated using the formula in Definition 17.

4.3 Experiment 1.2: Global Structure vs. GNN Performance for Passing and Failing Nodes

Goal. To conduct a further study we dive deeper and examine how correctly classified (*Passing*) and misclassified (*Failing*) test nodes differ from global graph statistics, focusing exclusively on homophily given its strong correlation with accuracy (Table 3).

Method. Using the definitions in Appendix B, we compare the average local homophily of each node group to the global homophily of the graph. Let η_{Global} denote the global homophily and $\eta_{\text{local}}(v)$ the per-node local homophily (Definition 5). For a node group $S \subseteq V$, such as the correctly classified $\mathcal{P}$ the misclassified $\mathcal{F}$ test nodes (Definition 6), we compute its mean local homophily $\bar{\eta}_S = \frac{1}{|S|} \sum_{v \in S} \eta_{\text{local}}(v)$.

The deviation from the global homophily level is then $\Delta_{\text{Group–Global}}(h) = \bar{\eta}_S - \eta_{\text{Global}}, S \in \{\mathcal{P}, \mathcal{F}\}$.

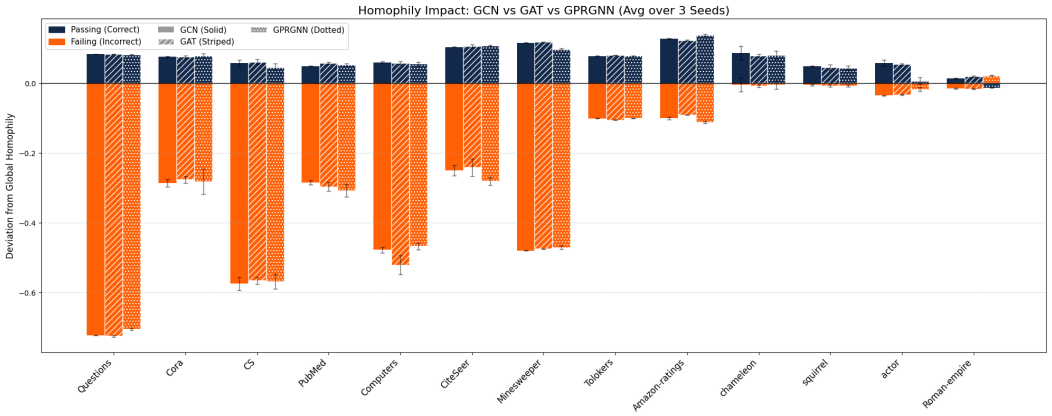


Figure 1: $\Delta_{\text{Passing–Global}}(\eta)$ and $\Delta_{\text{Failing–Global}}(\eta)$ for GCN, GAT, and GPRGNN models

Results. As shown in Figure 1, Passing nodes exhibit mainly positive *Passing–Global* differences in homophily (e.g., 0.073 for Cora GCN, 0.075 for Cora GAT), indicating that their neighborhoods are at least as homophilic as the global graph. In contrast, Failing nodes show consistently negative *Failing–Global* differences (e.g., -0.279 for Cora GCN, -0.2917 for Cora GAT and -0.723 for Questions GCN), meaning their local homophily is substantially lower than the global homophily. The magnitude of these negative deviations is much larger than the positive deviations for passing nodes, confirming that low homophily is a strong predictor of misclassification and that heterogeneous neighborhoods provide limited useful signal for message passing. Notably, the use of a GPRGNN, which was specially designed to deal with heterophilous datasets, does not modify significantly the differences observed, except for the Roman-empire dataset, where sign of the difference flipped.

4.4 Experiment 2.1: Comparing GNNs to Majority-Vote Baselines

Goal. Prior experiments showed that nodes with higher homophily achieve higher accuracy. Here we isolate how much of this predictive power comes from simple label agreement by comparing a Node-Label Majority Baseline (NLM) with three message-passing models (GCN, GAT, and GPRGNN). This tests whether performance in homophilic graphs reflects meaningful message passing or merely local label homogeneity.

Method. For each dataset, we compute the accuracy of NLM, which assigns each node the most frequent label among its 1-hop neighbors. We then compare these results to the performance of GCN and GAT.

Results. Figure 2 reports accuracy for the NLM, GCN, GAT, and GPRGNN models across all datasets. We observe that NLM either outperforms or matches GCN on 1/3 of all datasets (primarily with high global homophily). The only dataset where we can see a significant gap in performance for NLM is the Roman-Empire dataset, which exhibits the lowest homophily (Table 2). This pattern indicates that in strongly homophilic graphs (e.g., Cora, CiteSeer, PubMed), local label agreement alone explains a large portion of GNN predictive performance. In contrast, when homophily is weak or inverted, the predictive signal from neighborhood labels becomes unreliable and message passing provides additional value by aggregating feature and structural cues beyond local label agreement. Overall, these findings challenge the assumption that message passing is the main driver of GNN performance in homophilic settings. Much of the observed accuracy in such graphs can be attributed to the inherent correlation of labels in local neighborhoods, with message passing adding only marginal improvements. However, there are notable exceptions (e.g., Questions, Coauthor-CS), where homophily is relatively high yet NLM underperforms GCN. These cases suggest that message passing can still provide additional predictive benefit beyond label agreement, particularly when node features carry complementary information.

4.5 Experiment 2.2: Comparing GNNs to Feature-Only MLP Baselines

Goal. Previous experiments showed that relying solely on label agreement yields performance comparable to GNNs on highly homophilous graphs. However, there are cases where message-passing models still outperform the NLM baseline, suggesting that additional information contained

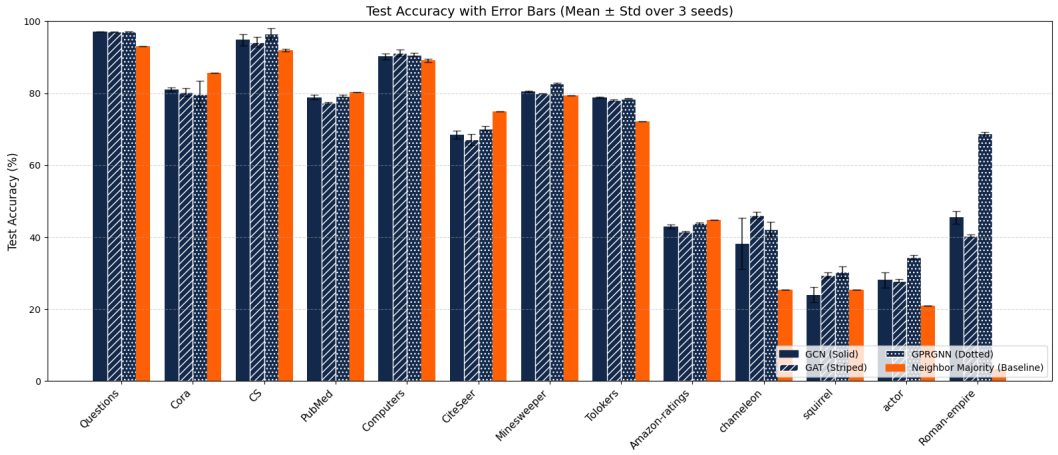


Figure 2: **Accuracy comparison between a homophily-based Neighbor Majority baseline (NLM) and GNNs (GCN, GAT, and GPRGNN).** Bars show test accuracy across all datasets for GCN (blue), GAT (striped blue), GPRGNN (dotted blue), and the NLM baseline (orange). The NLM outperforms or matches GNN performance on many homophilous datasets (e.g., Cora, PubMed, CiteSeer), indicating that local label agreement alone can account for a substantial portion of predictive accuracy. In contrast, on heterophilous datasets (e.g., Chameleon, Actor), the NLM fails while GCN and GAT maintain moderate accuracy, indicating that message passing becomes essential when homophily provides unreliable or misleading label signals. We see significant improvements in the accuracy provided by the GPRGNN model for the actor and Roman empire datasets.

in the node embeddings contributes to their predictive power. To further disentangle the contribution of message passing from feature-based learning, we compare GCN, GAT, and GPRGNN models with a Multi-Layer Perceptron (MLP) (Definition 8) that operates only on node features. This experiment isolates how much predictive signal arises from intrinsic node features versus relational structure.

Method. We train a two-layer MLP on the raw node features and evaluate its accuracy on the same train/validation/test splits used for the GNNs. The MLP has no access to graph connectivity, while GNNs leverages neighborhood information through message passing and aggregation.

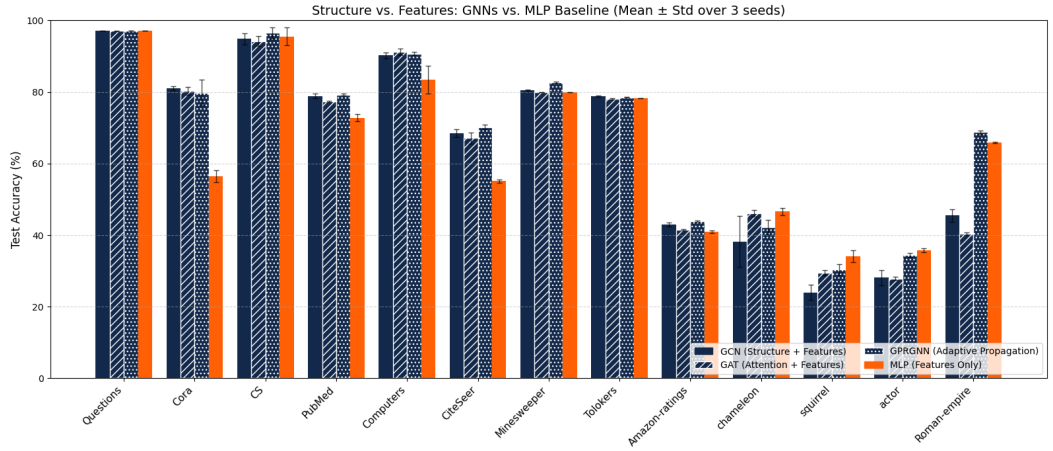
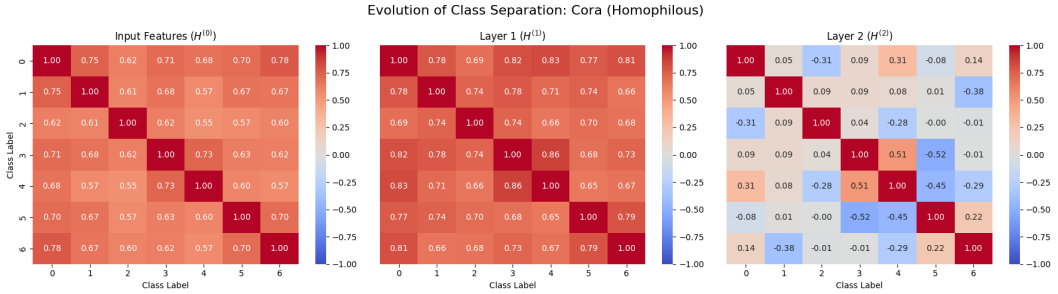
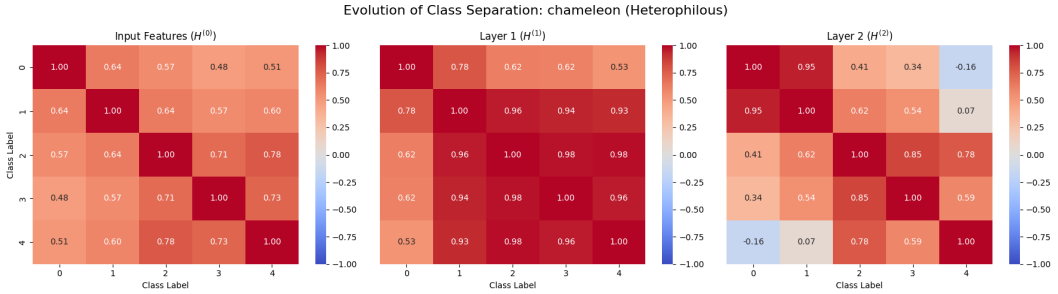


Figure 3: **Accuracy comparison between feature-only MLPs and message-passing GNNs across all datasets.** Bars show test accuracy for three models: a feature-only MLP (orange), a GCN (dark blue), a GAT (striped blue), and a GPRGNN(dotted blue). The results illustrate how much predictive signal comes from node features alone versus message passing.

Results. Figure 3 summarizes test accuracy for the feature-only MLP, GCN, GAT, and GPRGNN across all datasets. A clear pattern emerges. On strongly homophilous datasets such as Cora,



(a) High-homophily dataset



(b) Heterophilous dataset

Figure 4: **Class-wise cosine similarity of averaged node embeddings after message passing in GCN.** Each heatmap visualizes the similarity matrix $\mathbf{S}_{ij} = \mathrm{Sim}(c_i, c_j)$ between class-averaged embeddings (Definitions 14, 15).

PubMed, Computers, and CiteSeer, all GNN models substantially outperform the MLP, demonstrating that message passing provides significant predictive benefit when structural information aligns with labels. Interestingly, the Questions dataset shows high performance across all models, including the feature-only MLP. Although the graph exhibits strong homophily, the near-identical performance of GNNs and MLP indicates that node features alone carry sufficient discriminative signal, and message passing contributes only marginal gains.
On heterophilous datasets such as Chameleon, Squirrel, and Actor, the MLP matches or even exceeds the performance of all GNNs. This indicates that when neighborhood labels do not correlate with the target label, message passing introduces noise rather than useful signal, and feature-based learning becomes more reliable. Overall, the figure highlights a consistent structural dependence: message passing is beneficial primarily in high-homophily regimes, while in heterophilous settings, node features alone often provide a stronger and more stable predictive signal than graph-based aggregation.

4.6 Experiment 3.1: Class-wise Similarity After Message Passing

Goal. Previous experiments showed that message passing enhances the representational power of node embeddings by incorporating information from their neighborhoods, beyond what is contained in the raw features alone. In this section, we investigate how each transformation inside a GNN layer reshapes node representations. Specifically, we measure how class-wise averaged embeddings separate after each layer, and whether message passing collapses distinct classes or enhances inter-class separation. Due to time constrains we were able to perform this study only for GCN model.

Method. To analyze how the graph topology progressively alters the feature space, we compute the class-wise averaged embeddings $\bar{h}_c^{(\ell)}$ (Definition 14) at three specific depths: the raw input features ($\ell = 0$), the first hidden layer ($\ell = 1$), and the final output ($\ell = 2$). For each depth $\ell \in \{0, 1, 2\}$, we compute the pairwise cosine similarity matrix:

$$\mathbf{S}_{ij}^{(\ell)} = \mathrm{Sim}(\bar{h}_{c_i}^{(\ell)}, \bar{h}_{c_j}^{(\ell)}),$$

and visualize it as a heatmap to observe whether the message-passing mechanism disentangles or mixes the class representations.

Results. Figure 4 shows that message passing produces clear class-level geometry: high-homophily datasets exhibit strong block-diagonal structure, indicating tight within-class clustering and low cross-class similarity. In contrast, heterophilous datasets show weaker separation: many classes display high off-diagonal similarity, suggesting that message passing inadvertently mixes representations across class boundaries. Overall, the results indicate that message passing sharpens class structure when homophily is high, but struggles to maintain class separation in heterophilous graphs, providing further evidence of the structural dependence observed in Sections 4.4 - 4.5.

4.7 Experiment 3.2: Message Passing and Representation Dynamics

Goal. Previous experiments showed that feature mixup occurs in heterophilous graphs, where message passing blends information from dissimilar neighbors. In this section, we investigate how message passing reshapes node embeddings and whether these representation shifts align with neighborhood homophily. Our goal is to determine whether successful classification corresponds to embeddings being pulled toward homophilous neighbors and, conversely, whether misclassification arises from weak or misaligned representation drift.

Method. Using the message-passing drift metric $\Delta(v)$ (Definition 13), we analyze how strongly each node's final embedding aligns with averaged homophilous versus heterophilous neighbor message vectors. We partition nodes into correctly and incorrectly classified sets,

$$\mathcal{P} = \{v : \hat{y}_v = y_v\}, \qquad \mathcal{F} = \{v : \hat{y}_v \neq y_v\},$$

as introduced in Definition 6. Then we plot the average drift metric $\Delta(v)$ for all the nodes in $\mathcal{P}$ and $\mathcal{F}$.

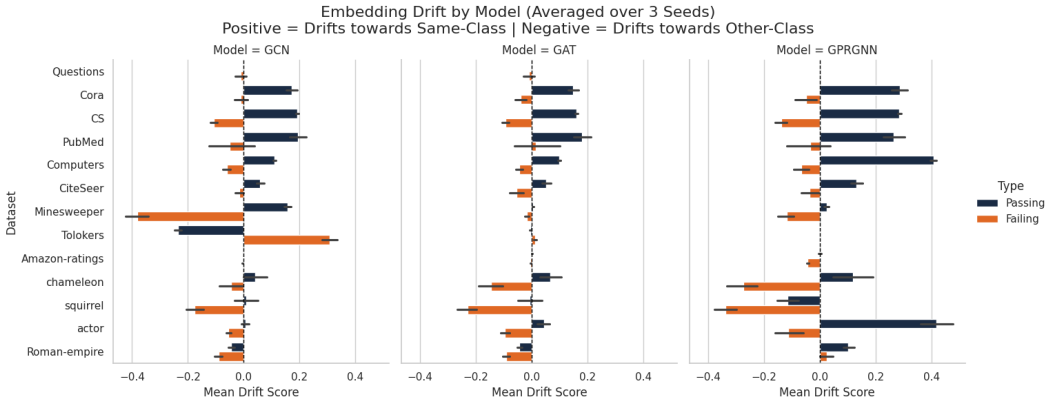


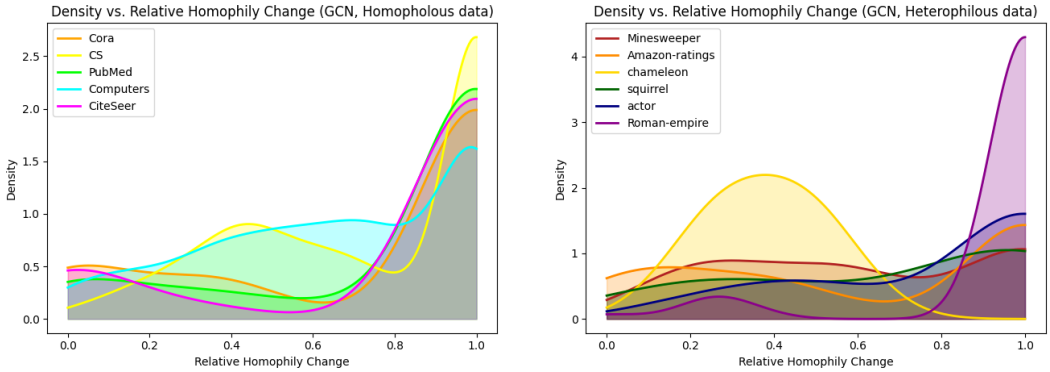
Figure 5: **Message-passing drift ($\Delta(v)$) across architectures.** Average embedding drift for passing and failing test nodes in GCN, GAT, and GPRGNN. Results are averaged over three random seeds, with 95% confidence intervals shown. Positive values indicate drift towards same-class neighbors (homophily), while negative values indicate drift towards other-class neighbors (heterophily).

Results. Results are shown in Figure 5. A positive drift value $\Delta(v) > 0$ indicates that the final embedding of v is more aligned with messages from homophilous neighbors, whereas a negative value $\Delta(v) < 0$ reflects stronger alignment with heterophilous neighbors. Across all datasets—except for a few strongly heterophilous ones—passing nodes exhibit predominantly positive drift, meaning their final representations move closer to the centroids of homophilous neighbors. In contrast, failing nodes show mostly negative drift, indicating that their embeddings are pulled toward heterophilous neighborhoods, consistent with misaligned or uninformative message-passing dynamics. We also conducted statistical testing in Section E.

4.8 Experiment 4: Robustness of Node Predictions Under Homophily Reduction

Goal. This experiment examines how dependent GNN predictions are on local homophily $\eta_{\text{local}}(v)$. We investigate: (i) how much homophily must be removed before a correctly classified node flips, (ii) how many nodes remain stable even after $\eta_{\text{local}}(v)$ is reduced to zero.

Method. For each correctly classified node $v \in \mathcal{P}$, we apply a homophily-reduction perturbation by iteratively removing homophilous neighbors and recomputing prediction $\hat{y}_v$ at each step. For every node, we track (i) the node's original homophily $\eta^0(v)$ before any pertubations are made, (ii)



(a) Homophilous datasets (e.g., Cora, Citeseer, PubMed)

(b) Heterophilous datasets (e.g., Chameleon, Squirrel, Actor, Minesweeper)

Figure 6: **Sensitivity of GCN predictions to homophily reduction.** Left: probability density curves of $\frac{\Delta\eta(v)}{\eta^0(v)}$ for homophilous datasets. Right: histograms for heterophilous datasets. Large required reductions indicate strong reliance on homophilous neighborhoods. Small or near-zero reductions reflect limited dependence on homophily and substantially higher fragility to perturbations.

how many homophilous neighbors must be removed before the prediction flips, (iii) whether the prediction ever flips at all, (iv) the resulting homophily drop $\Delta\eta(v)$ required to induce misclassification (as defined in Definition 16). We then analyze the distribution of the relative homophily change, defined as $\Delta\eta(v)$ divided by $\eta^0(v)$, as a probability density function (converted using the formula in Definition 18) across the nodes that flip, as well as the proportion of nodes that remain correctly classified even when their local homophily is reduced to zero.

Results. Figures 6 and 7 show the distributions of the relative homophily change required to cause misclassifications for GCN and GAT. Due to time constraints, we weren't able to perform these experiments for the third model. Both architectures exhibit a similar pattern of requiring large reductions in local homophily, more often reducing it to zero, before a node is misclassified. This holds for both homophilous (Cora, Citeseer, PubMed) and heterophilous (Chameleon, Squirrel, Actor, Minesweeper) datasets, showing that local homophily may be the dominant factor in node classification.
The stability analysis in Figure 8 reveals a clear architectural contrast across homophilous and heterophilous regimes. GCN maintains a higher fraction of unchanged predictions on homophilous datasets, consistent with its reliance on uniform neighborhood averaging when most neighbors provide aligned, informative signals. In contrast, GAT exhibits higher stability on heterophilous datasets, suggesting that its attention mechanism can better isolate the minority of useful neighbors in label-inconsistent or noisy neighborhoods.

5 Conclusion

Together, the results of Experiments 1–4 reveal a coherent picture of how homophily shapes GNN behavior. Experiment 1 showed that correctly classified nodes tend to inhabit high-homophily, well-connected neighborhoods, while misclassified nodes are typically situated in structurally irregular or low-homophily regions. Experiment 2 further demonstrated that much of the predictive performance in homophilic graphs can be explained by simple label agreement: a majority-vote classifier performed comparably or outperformed GNNs, and feature-only MLPs perform substantially worse, highlighting the critical role of neighborhood structure. Experiment 3 then provides a mechanism-level explanation for these patterns. The message-passing drift $\Delta(v)$ quantifies how strongly a node's representation is pulled toward homophilous or heterophilous neighbors, and we observe a clear alignment between positive drift and correct classification. Passing nodes show consistently positive drift, reflecting effective propagation of label-consistent signals, whereas Failing nodes exhibit weak or negative drift, indicating that message passing amplifies noisy or conflicting neighborhood information. These findings collectively suggest that GNN performance is tightly coupled to the direction and strength of representation drift induced by homophily, and that failures occur when message passing is unable to overcome local structural or label noise.

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 *International conference on machine learning*, 2019.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. *International Conference on Learning Representations*, 2021.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pp. 1407–1418. PMLR, 2021.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=n6jl7fLxrP.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Qimai Li, Zhichao Han, and Xiaoming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI conference on artificial intelligence*, 2018.

Donald Loveland, Jiong Zhu, Mark Heimann, Benjamin Fish, Michael T. Schaub, and Danai Koutra. On performance discrepancies across local homophily levels in graph neural networks. In *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pp. 6:1–6:30. PMLR, 2024. URL https://arxiv.org/pdf/2306.05557.

Kangkang Lu, Yanhua Yu, Zhiyong Huang, and Tat-Seng Chua. Enhancing spectral graph neural networks with llm-predicted homophily. *arXiv preprint arXiv:2506.14220*, 2025.

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 impact of homophily principle on node distinguishability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Ana Lucic, Maartje ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. *arXiv preprint arXiv:2102.03322*, 2021. doi: 10.48550/arXiv.2102.03322.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19620–19631, 2020.

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

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, volume 8, pp. 1, 2012.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

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 (ICLR)*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

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? *arXiv preprint arXiv:2302.11640*, 2023.

Pavel Procházka, Michal Mareš, and Marek Dědič. Which graph properties affect gnn performance for a given downstream task? In *23rd Conference Information Technologies – Applications and Theory*, 2023. URL https://ceur-ws.org/Vol-3498/paper7.pdf.

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

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–93, 2008.

Oleksandr Shchur, Matthias Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop, NeurIPS*, 2018.

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

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. PMLR, 2019.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Xin Zheng, Yi Wang, Yixin Liu, Ming Li, Miao Zhang, Di Jin, Philip S Yu, and Shirui Pan. Graph neural networks for graphs with heterophily: A survey. *arXiv preprint arXiv:2202.07082*, 2022.

Yilun Zheng, Sitao Luan, and Lihui Chen. What is missing in homophily? disentangling graph homophily for graph neural networks. *arXiv preprint arXiv:2406.18854*, 2024.

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, 2020.

A SUMMARY OF METRICS USED

Table 4: Global and local graph statistics used in our analysis. We denote a graph by $G = (V, E, X)$ with $|V| = n$ nodes, $|E| = m$ edges, features $X \in \mathbb{R}^{n \times d}$ with $x_v \in X$, and labels $\{y_v\}_{v \in V}$.

Metric	Notation / Formula	Description
Global Statistics		
Homophily	$\eta = \frac{\lvert\{(u,v) \in E : y_u = y_v\}\rvert}{\lvert E \rvert}$	Fraction of edges whose endpoints share the same label. Captures global label agreement across the graph.
Average degree	$\bar{d} = \frac{1}{\lvert V \rvert} \sum_{v \in V} \deg(v)$	Mean number of neighbors per node. Measures global connectivity.
Global clustering coefficient	$C = \frac{3T}{W}$	Triangle density, where T is the number of triangles and W the number of connected triplets. Measures global triadic closure.
Attribute similarity	$s = \frac{1}{\lvert E \rvert} \sum_{(u,v) \in E} \frac{x_u^\top x_v}{\lVert x_u \rVert \lVert x_v \rVert}$	Average cosine similarity of node features across edges. Captures global feature-level homogeneity.
Local Statistics		
Per-node degree	$\deg(v) = \lvert \mathcal{N}(v) \rvert$	Number of 1-hop neighbors ($\mathcal{N}(v)$) of node v. Measures local connectivity.
Per-node clustering coefficient	$C(v) = \frac{2e_v}{k(k-1)}, \quad k = \deg(v)$	Fraction of possible e_v - edges among neighbors of v - that exist. Measures local triangle density.
Local homophily	$\eta_{\text{local}}(v) = \frac{1}{\max(\deg(v), 1)} \sum_{u \in \mathcal{N}(v)} \mathbb{I}[y_u = y_v]$	Proportion of neighbors of v that share the same label. Captures node-level label agreement.
Passing vs. failing nodes	$\mathcal{P} = \{v : \hat{y}_v = y_v\}, \quad \mathcal{F} = \{v : \hat{y}_v \neq y_v\}$	Groups nodes into correctly and incorrectly classified sets based on model predictions.

B DEFINITIONS

B.1 GRAPH STATISTICS

To characterize the graph, we extract statistics at the global (whole graph) and local (neighborhood) levels. Table 4 summarizes both sets of metrics and their formal definitions.

Definition 1 (Global Homophily)**.** *The global homophily of a graph is*

$$\eta_{Global} = \frac{|\{(u,v) \in E \ : \ y_u = y_v\}|}{|E|}.$$

This measures the fraction of edges connecting same-label nodes.

Definition 2 (Average Degree)**.** *The average degree is the mean number of neighbors per node:*

$$\bar{d} = \frac{1}{|V|} \sum_{v \in V} \deg(v),$$

where $\deg(v)$ is the degree of node v.

Definition 3 (Global Clustering Coefficient)**.** *The global clustering coefficient quantifies the density of triangles:*

$$C = \frac{3T}{W},$$

where T is the total number of triangles and W is the number of connected triplets.

Definition 4 (Attribute Similarity)**.** *Attribute similarity measures average cosine similarity of feature vectors across edges:*

$$s = \frac{1}{|E|} \sum_{(u,v) \in E} \frac{x_u^\top x_v}{\|x_u\| \, \|x_v\|}.$$

B.2 Local (Node-level) Statistics

To link node-level structural properties with predictive success, we define per-node quantities and a group-wise aggregation procedure.

Definition 5 (Per-node Local Homophily)**.** *Given the true label y_v of node v, we define its local homophily as:*

$$\eta_{local}(v) = \frac{1}{\max(\deg(v), 1)} \sum_{u \in \mathcal{N}(v)} \mathbb{I}[y_u = y_v].$$

This measures the proportion of v's neighbors that share its label.

Definition 6 (Passing and Failing Nodes)**.** *Given model predictions $\hat{y}_v$, we partition nodes into:*

$$\mathcal{P} = \{v \mid \hat{y}_v = y_v\}, \qquad \mathcal{F} = \{v \mid \hat{y}_v \neq y_v\},$$

representing correctly (passing) and incorrectly (failing) classified nodes, respectively.

B.3 Baseline Models

To isolate the effects of homophily and message passing, we compare GNNs to two non-message-passing baselines: a simple node-label majority classifier and a feature-only multilayer perceptron (MLP). We formally define both below.

Definition 7 (Node-Label Majority Baseline (NLM))**.** *For each node $v \in V$, let $\mathcal{N}(v)$ denote its 1-hop neighborhood and let y_u be the label of a neighbor u. Let $\mathcal{C}$ denote the set of possible class labels. The majority-vote classifier predicts the label of v as:*

$$\hat{y}_v = \arg\max_{c \in \mathcal{C}} \big|\{u \in \mathcal{N}(v) : y_u = c\}\big|.$$

In cases of ties, a random label among the tied classes is selected. This baseline uses only the labels of neighboring nodes and captures the predictive power attributable purely to local homophily.

Definition 8 (Feature-Only MLP Baseline)**.** *The feature-only MLP removes all structural information by operating solely on node attributes. Given node features $x_v \in \mathbb{R}^d$, the MLP computes:*

$$h_v^{(1)} = \sigma(W_1 x_v + b_1), \qquad h_v^{(2)} = \sigma(W_2 h_v^{(1)} + b_2), \qquad \hat{y}_v = \mathrm{softmax}(W_3 h_v^{(2)} + b_3),$$

where W_i, b_i are trainable parameters and $\sigma(\cdot)$ denotes a nonlinear activation function.

Because no graph connectivity information is used, this model quantifies how much predictive signal is contained purely in node features, independent of homophily or message passing.

These two baselines represent opposite extremes: the majority-vote classifier NLM depends entirely on neighborhood labels, while the MLP depends entirely on node features. GNN performance relative to these baselines helps isolate the contribution of homophily and message passing.

B.4 Representation Dynamics

In this section, we introduce a set of metrics designed to quantify how message passing reshapes node embeddings and how these representations interact with their local neighborhoods. Our goal is to characterize the directional influence of homophilous and heterophilous neighbors on a node's final embedding and to understand how these dynamics relate to classification success or failure.

Definition 9 (Message Passing)**.** *A message-passing GNN updates node representations by aggregating information from their neighbors. Given node features $\{h_v^{(k)}\}_{v\in V}$ at layer k, the update rule for node v at layer $k+1$ is*

$$h_v^{(k+1)} = \phi^{(k)}\Big(h_v^{(k)}, \square_{u\in\mathcal{N}(v)}\, \psi^{(k)}(h_v^{(k)}, h_u^{(k)})\Big),$$

where $\psi^{(k)}$ is a message function, $\square$ is a permutation-invariant aggregation operator (e.g., sum, mean, attention), and $\phi^{(k)}$ is an update function. This process iteratively mixes information from each node's neighborhood.

Definition 10 (GPRGNN)**.** *The GPRGNN architecture modifies the general structure of a GNN in an attempt to deal with heterophily. As a first step, it uses an learnable function f_θ (generally an MLP), to obtain initial logits as follows:*

$$H^{(0)} = f_\theta(X)$$

where X are the initial node features. Then, the model combines information from the graph structure through a series weighted hops as follows:

$$Z = \sum_{k=0}^{K} \gamma_k \hat{A}^k H^{(0)}$$

Where $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ is the symmetrically normalized adjacency matrix with self loops, $H^{(0)}$ are the initial logits, K is the depth of propagation, and γ_k are the learnable scalars for each hop. By tuning the values of γ_k at training, the GPRGNN architecture is able to increase, decrease or even have the neighborhood at a certain hop provide a negative singnal. This should allow the model deal better with close-neighbor heterophily.

Definition 11 (Averaged Messages from Homophilous and Heterophilous Neighbors)**.** *Given a node v with label y_v and neighborhood $\mathcal{N}(v)$, we partition neighbors into homophilous and heterophilous sets:*

$$\mathcal{N}^+(v) = \{u \in \mathcal{N}(v) : y_u = y_v\}, \qquad \mathcal{N}^-(v) = \{u \in \mathcal{N}(v) : y_u \neq y_v\}.$$

For a message-passing GNN with message function $\psi^{(k)}$, the aggregated messages received by v from these two groups at layer k are defined as:

$$\bar{m}_v^{+(k)} = \frac{1}{\max(|\mathcal{N}^+(v)|, 1)} \sum_{u\in\mathcal{N}^+(v)} \psi^{(k)}\Big(h_v^{(k)}, h_u^{(k)}\Big),$$

$$\bar{m}_v^{-(k)} = \frac{1}{\max(|\mathcal{N}^-(v)|, 1)} \sum_{u\in\mathcal{N}^-(v)} \psi^{(k)}\Big(h_v^{(k)}, h_u^{(k)}\Big).$$

These represent the mean incoming messages to v from homophilous and heterophilous neighbors, respectively, before the update function is applied.

Definition 12 (Cosine Similarity to Homophilous and Heterophilous Message Averages)**.** *Let $h_v^{(K+1)}$ be the final embedding of node v after K message-passing layers, and let $\bar{m}_v^{+(K)}$ and $\bar{m}_v^{-(K)}$ denote the averaged incoming messages from homophilous and heterophilous neighbors at layer K (Definition 11). The cosine similarity between the node's embedding and each averaged message vector is defined as:*

$$\cos^+(v) = \frac{h_v^{(K+1)\top}\, \bar{m}_v^{+(K)}}{\|h_v^{(K+1)}\|\, \|\bar{m}_v^{+(K)}\|}, \qquad \cos^-(v) = \frac{h_v^{(K+1)\top}\, \bar{m}_v^{-(K)}}{\|h_v^{(K+1)}\|\, \|\bar{m}_v^{-(K)}\|}.$$

A higher $\cos^+(v)$ indicates that the final node embedding is more aligned with messages from homophilous neighbors, while a higher $\cos^-(v)$ indicates stronger influence from heterophilous neighbors.

Definition 13 (Message-Passing Drift)**.** *Let* $\cos^+(v)$ *and* $\cos^-(v)$ *denote the cosine similarity between the final node embedding* $h_v^{(K)}$ *and the averaged messages received from homophilous and heterophilous neighbors, respectively (Definition 12). We define the* message-passing drift *of node* v *as:*

$$\Delta(v) = \cos^+(v) - \cos^-(v).$$

A positive value $\Delta(v) > 0$ indicates that the final embedding of v is more aligned with messages from homophilous neighbors, while a negative value $\Delta(v) < 0$ indicates stronger alignment with heterophilous neighbors. This provides a directional measure of how message passing pulls a node's representation within the embedding space.

Definition 14 (Class-wise Embedding)**.** *Let* $\mathcal{C}$ *be the set of classes and* $H^{(\ell)} \in \mathbb{R}^{N\times d}$ *be the node embedding matrix at layer* ℓ*. We define the base case* $H^{(0)} = X$ *as the raw input features. For any layer* $\ell \geq 0$*, the class-wise averaged embedding for class* $c \in \mathcal{C}$ *is defined as:*

$$\bar{h}_c^{(\ell)} = \frac{1}{|\mathcal{V}_c|} \sum_{v\in\mathcal{V}_c} h_v^{(\ell)}, \qquad \mathcal{V}_c = \{v \in V : y_v = c\},$$

where $h_v^{(\ell)}$ *denotes the row vector of node* v *in* $H^{(\ell)}$*.*

Definition 15 (Class-wise Cosine Similarity)**.** *Given two classes* $c_i, c_j \in \mathcal{C}$ *with averaged embeddings* $\bar{h}_{c_i}$ *and* $\bar{h}_{c_j}$*, their class-wise similarity is*

$$\mathrm{Sim}(c_i, c_j) = \frac{\bar{h}_{c_i}^\top \bar{h}_{c_j}}{\|\bar{h}_{c_i}\| \, \|\bar{h}_{c_j}\|}.$$

Large values indicate that the two classes become aligned in the GNN embedding space, whereas small or negative values indicate strong inter-class separation.

Definition 16 (Homophily-Reduction Perturbation)**.** *For a correctly classified node* $v \in \mathcal{P}$*, we define a homophily-reduction perturbation as the iterative removal of edges* (v, u) *where* u *is homophilous with* v *(i.e.,* $y_u = y_v$*). After each removal, the model's predicted label* $\hat{y}_v$ *is recomputed. Let*

$$\Delta\eta(v) = \eta_{\textit{local}}^{\textit{before}}(v) - \eta_{\textit{local}}^{\textit{after}}(v)$$

denote the decrease in local homophily caused by the perturbation. The process terminates when v *becomes misclassified (i.e.,* $\hat{y}_v \neq y_v$*,) or when all homophilous neighbors have been removed.*

Definition 17 (Pearson Correlation)**.** *Given two real-valued sequences* $\{a_i\}_{i=1}^n$ *and* $\{b_i\}_{i=1}^n$*, the Pearson correlation coefficient measures the strength of their linear relationship and is defined as:*

$$\rho(a, b) = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \, \sqrt{\sum_{i=1}^n (b_i - \bar{b})^2}},$$

where $\bar{a}$ *and* $\bar{b}$ *denote the sample means of the two sequences.*

A value of $\rho(a, b) = 1$ indicates perfect positive linear correlation, -1 indicates perfect negative linear correlation, and 0 indicates no linear correlation.

Definition 18 (Kernel Density Estimation)**.** *Given a distribution* X *made up of variables* $\{X_1, X_2, \ldots, X_n\}$*, a curve is defined by a function*

$$f(x) = \frac{1}{(n \cdot h)} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{n}\Big)$$

where K *is the Gaussian Kernel function, and* h *is a smoothing parameter that is determined by Scott's Rule (*$h = n^{-\frac{1}{d+4}}$*), where* d *the dimension of the data (in this case* $d = 1$*).*

C DATASETS

- **Questions** (Platonov et al., 2023): A question–answer interaction graph derived from a community Q&A platform. Nodes represent questions, and edges connect questions answered by the same user. The task is to classify questions into topical categories based on textual content and interaction patterns. Node features are 300-dimensional embeddings computed from question text. The dataset includes 10 predefined train/validation/test splits. **Statistics:** 48,921 nodes; 153,540 edges; 10 classes; 300 features.

- **Cora** (Sen et al., 2008): A citation network of machine learning papers, consisting of 2,708 nodes, 5,429 edges, and 7 classes. It exhibits high homophily and is commonly used to benchmark message-passing GNNs.
- **Coauthor-CS** (Shchur et al., 2018): A co-authorship network of computer science researchers, containing 18,333 nodes, 81,894 edges, and 15 classes. The graph is highly homophilic and features rich continuous attributes, providing contrast to citation networks.
- **PubMed** (Namata et al., 2012): A larger biomedical citation network with 19,717 nodes, 44,338 edges, and 3 classes. PubMed has moderately high homophily and more challenging class imbalance patterns.
- **Amazon-Computers** (Shchur et al., 2018): A co-purchase graph of electronic products from Amazon, with 13,381 nodes, 245,778 edges, and 10 classes. Amazon-Computers exhibits moderately low homophily and is often used to evaluate robustness under weaker structural label correlations.
- **CiteSeer** (Sen et al., 2008): A citation graph of computer science publications with 3,327 nodes, 4,732 edges, and 6 classes.
- **Minesweeper** (Platonov et al., 2023): A synthetic graph inspired by the Minesweeper game. The graph is a 100×100 grid where each node (cell) connects to up to eight neighbors, with fewer connections along the boundaries. Twenty percent of nodes are randomly designated as mines, and the task is to classify nodes as mine vs. safe. Node features encode the number of neighboring mines using a 7-dimensional one-hot vector; for 50% of nodes, features are masked as unknown using an additional binary indicator. **Statistics:** 10,000 nodes; 78,804 edges; 2 classes; 7 features.
- **Tolokers** (Platonov et al., 2023): A crowdsourcing interaction network constructed from the Toloka platform. Nodes represent workers ("tolokers"), and edges connect workers who have participated in the same task. The prediction task is to identify workers who were banned in a specific project. Node features summarize worker profile attributes and task performance statistics. The dataset is moderately large and highly connected. **Statistics:** 11,758 nodes; 1,038,000 edges; 2 classes; 10 features.
- **Amazon-ratings** (Platonov et al., 2023): A heterophilous product co-purchasing graph based on Amazon metadata. Nodes represent products (e.g., books, CDs, DVDs), and edges connect items frequently bought together. The prediction task is to classify products into five rating categories derived from user review averages. Node features are 300-dimensional embeddings obtained by averaging word vectors from the product descriptions. **Statistics:** 24,492 nodes; 186,100 edges; 5 classes; 300 features.
- **Chameleon** (Rozemberczki et al., 2021; Pei et al., 2020): A Wikipedia page–page network containing articles related to the Chameleon web template engine and similar topics. Nodes represent articles, and edges represent mutual hyperlinks between them. The graph is characterized by high heterophily, as connected articles frequently belong to different categories. The prediction task is to classify articles into five discrete popularity levels based on their average monthly traffic. Node features are bag-of-words vectors indicating the presence of informative nouns in the article text. **Statistics:** 2,277 nodes; 36,101 edges; 5 classes; 2,325 features.
- **Squirrel** (Pei et al., 2020): A Wikipedia page–page network constructed similarly to Chameleon, but focusing on articles related to squirrels. Nodes are pages and edges are mutual links, resulting in a graph with very high density and strong heterophily. The task is to predict the average monthly traffic level of the page (5 classes). Like Chameleon, the node features are bag-of-words representations of the article content. It is often considered one of the most challenging benchmarks for standard GNNs due to its complex structural patterns. **Statistics:** 5,201 nodes; 217,073 edges; 5 classes; 2,089 features.
- **Actor** (Pei et al., 2020): An actor co-occurrence network derived from a larger film-director-actor-writer graph. Nodes represent actors, and edges connect actors who have appeared in the same film together. The prediction task involves classifying actors into five categories based on keywords from their Wikipedia pages (e.g., their primary genre or role types). The graph exhibits low homophily, meaning actors linked by co-occurrence often fall into different descriptive categories. Node features correspond to bag-of-words vectors

extracted from the actor's biographical information. **Statistics:** 7,600 nodes; 33,544 edges; 5 classes; 931 features.

- **Roman-empire** (Platonov et al., 2023): A heterophilous word-level graph constructed from the "Roman Empire" Wikipedia article. Nodes correspond to words and edges connect consecutive words or syntactically related words via dependency-tree links. Node features are 300-dimensional word embeddings, and node labels represent syntactic roles (17 frequent roles plus one aggregated class). **Statistics:** 22,662 nodes; 65,854 edges; 18 classes; 300 features.

D Hyperparameter Configurations

The hyperparameter configuration used for training the models is shown in the following table:

Table 5: Hyperparameter configuration for the graph neural network models.

Model	Hidden Dim.	Dropout	Model-Specific Parameters
GCN	64	0.5	–
GAT	64	0.6	Heads $= 8$
GPRGNN	64	0.5	$K = 10$, $\alpha = 0.1$, Init = PPR

We didn't observe meaningful changes in the performance by increasing either the depth of the models or the number of attention heads in the GAT model.

E Drift Analysis: Experimental Setup and Statistical Testing

Table 6: One-sided statistical tests evaluating the message-passing drift $\Delta(v)$ (Definition 13) for GCN and GAT across all test nodes. H1 tests whether Passing nodes (P) exhibit positive drift ($\Delta(v) > 0$); H2 tests whether Failing nodes (F) exhibit negative drift ($\Delta(v) < 0$); H3 tests whether Passing drift exceeds Failing drift ($\Delta_P > \Delta_F$). All tests use significance level $\alpha = 0.05$.

Model	Hypothesis	Mean Drift	Statistical Tests (p-values)	Decision
GCN	H1: $\Delta_P > 0$	0.03593 ($\lvert P\rvert = 12{,}371$)	t-test: 2.06×10^{-42}; Wilcoxon: 0	PASS
	H2: $\Delta_F < 0$	-0.02505 ($\lvert F\rvert = 5{,}229$)	t-test: 1.33×10^{-9}; Wilcoxon: 4.10×10^{-55}	PASS
	H3: $\Delta_P > \Delta_F$	$0.03593 > -0.02505$	Mann–Whitney U: 3.80×10^{-258}; Welch t-test: 8.29×10^{-35}	PASS
GAT	H1: $\Delta_P > 0$	0.03343 ($\lvert P\rvert = 12{,}336$)	t-test: 7.95×10^{-148}; Wilcoxon: 7.09×10^{-203}	PASS
	H2: $\Delta_F < 0$	-0.02604 ($\lvert F\rvert = 5{,}264$)	t-test: 3.88×10^{-34}; Wilcoxon: 4.48×10^{-35}	PASS
	H3: $\Delta_P > \Delta_F$	$0.03343 > -0.02604$	Mann–Whitney U: 2.89×10^{-144}; Welch t-test: 3.53×10^{-123}	PASS

E.1 Experimental Setup

We evaluate message-passing drift using two representative architectures (GCN and GAT) across a broad collection of homophilous and heterophilous node-classification benchmarks. For each dataset, models are trained using the standard `train/val/test` splits provided by PyG, and all statistical analyses are conducted on the held-out test nodes. Following training, we compute the drift $\Delta(v)$ for every node v, as defined in Definition 13.

We evaluate three directional hypotheses governing how message passing affects these two groups.

E.2 Statistical Hypotheses

We perform one-sided hypothesis tests at significance level $\alpha = 0.05$.

Hypothesis 1 (Passing nodes exhibit positive drift). For $v \in P$, we test

$$H_1^{(+)} : \mathbb{E}[\Delta(v) \mid v \in P] > 0,$$

using both a one-sample t-test and a Wilcoxon signed-rank test. These tests assess whether message passing systematically increases alignment with homophilous neighbors for correctly classified nodes.

Hypothesis 2 (Failing nodes exhibit negative drift). For $v \in F$, we test

$$H_1^{(-)} : \mathbb{E}[\Delta(v) \mid v \in F] < 0,$$

again using the t-test and Wilcoxon test. This evaluates whether message passing tends to pull misclassified nodes toward heterophilous neighborhoods.

Hypothesis 3 (Passing drift exceeds Failing drift). To compare the two groups directly, we test

$$H_1^{(\Delta)} : \mathbb{E}[\Delta(v) \mid v \in P] > \mathbb{E}[\Delta(v) \mid v \in F],$$

using the Mann–Whitney U test (non-parametric) and Welch's t-test (unequal variance). These tests quantify whether message passing produces a directional separation between representations of correct and incorrect nodes.

We briefly formalize the statistical tests used in our analysis.

One-sample t-test. Given observations $\{x_i\}_{i=1}^n$ with sample mean $\bar{x}$ and sample standard deviation s, the one-sample t-test evaluates the null hypothesis $H_0 : \mu = \mu_0$ by computing

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

which, under H_0, follows a Student-t distribution with $n-1$ degrees of freedom. We use $\mu_0 = 0$ to test whether mean drift differs from zero.

Wilcoxon signed-rank test. This is a non-parametric paired test that evaluates the null hypothesis $H_0 : \text{median}(x) = 0$ without assuming normality. Let $d_i = x_i - 0$ denote signed differences. The test ranks the absolute values $|d_i|$, assigns the corresponding signs, and computes the signed-rank statistic

$$W = \sum_{i=1}^{n} \text{sign}(d_i)\, \text{rank}(|d_i|).$$

Under H_0, W has a known distribution that yields exact or asymptotic p-values. This test detects whether the distribution of drift is shifted positively or negatively.

Mann–Whitney U test. Given two independent samples $X = \{x_i\}_{i=1}^{n_1}$ and $Y = \{y_j\}_{j=1}^{n_2}$, this non-parametric test evaluates whether X tends to take larger values than Y. Let R_i be the rank of each observation in the pooled sample. The U statistic is

$$U = \sum_{i=1}^{n_1} R_i - \frac{n_1(n_1+1)}{2},$$

which, under the null hypothesis $H_0 : X$ and Y are drawn from the same distribution, has a known sampling distribution. This test is sensitive to stochastic dominance rather than differences in means.

Welch's t-test. For two independent samples X and Y, Welch's t-test evaluates

$$H_0 : \mu_X = \mu_Y$$

without assuming equal variances. With sample means $\bar{x}, \bar{y}$ and sample variances s_X^2, s_Y^2, the test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_X^2/n_1 + s_Y^2/n_2}},$$

which follows a t-distribution with degrees of freedom determined by the Welch–Satterthwaite equation. This test accounts for unequal group sizes and heteroscedasticity, making it appropriate for comparing drift between the Passing and Failing sets.

E.3 Results

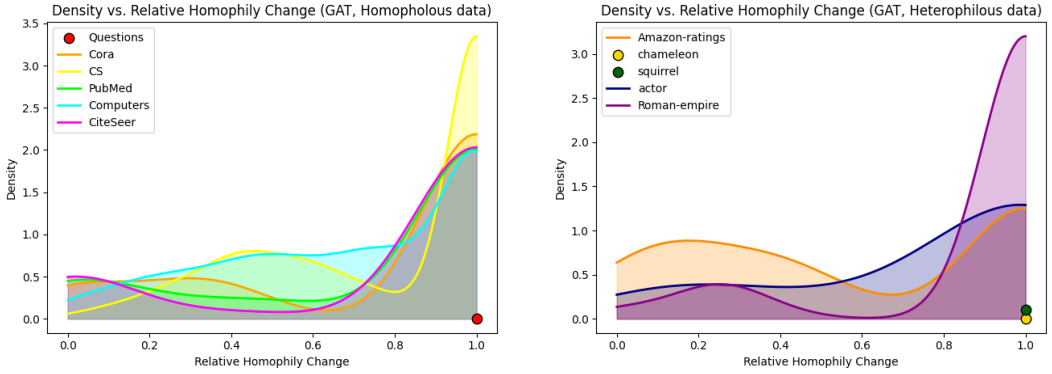
Table 6 reports mean drift values, associated p-values, and hypothesis decisions. Across all datasets and both architectures, we observe highly consistent directional drift behavior.

- **Passing nodes exhibit strongly positive drift.** For both GCN and GAT, one-sample tests yield extremely small p-values ($p < 10^{-42}$ for GCN; $p < 10^{-147}$ for GAT), indicating that message passing reliably aligns correct nodes toward homophilous neighbors.
- **Failing nodes exhibit significantly negative drift.** Failing nodes show strong negative drift (GCN: $p < 10^{-9}$; GAT: $p < 10^{-34}$), revealing that misclassified nodes are pulled toward heterophilous neighbors.
- **Passing drift significantly exceeds Failing drift.** Both Mann–Whitney and Welch tests reject the null with overwhelming significance (GCN: $p < 10^{-258}$; GAT: $p < 10^{-144}$), demonstrating that message passing induces a robust directional separation between P and F.

E.4 Conclusion

Taken together, these findings expose a consistent mechanism underlying message-passing GNN behavior: message passing *amplifies* homophilous signals for nodes whose labels are predicted correctly, while *propagating* heterophilous noise for nodes whose predictions are incorrect. Thus, both the successes and failures of GNNs can be interpreted through the directional effects of aggregation on neighborhood structure.

F Perturbation Experiment Extended



(a) Homophilous datasets (e.g., Cora, Citeseer, PubMed)

(b) Heterophilous datasets (e.g., Chameleon, Squirrel, Actor, Minesweeper)

Figure 7: **Sensitivity of GAT predictions to homophily reduction.** Left: probability density curves of $\frac{\Delta\eta(v)}{\eta^0(v)}$ for homophilous datasets. Right: histograms for heterophilous datasets. Large required reductions indicate strong reliance on homophilous neighborhoods. Small or near-zero reductions reflect limited dependence on homophily and substantially higher fragility to perturbations.

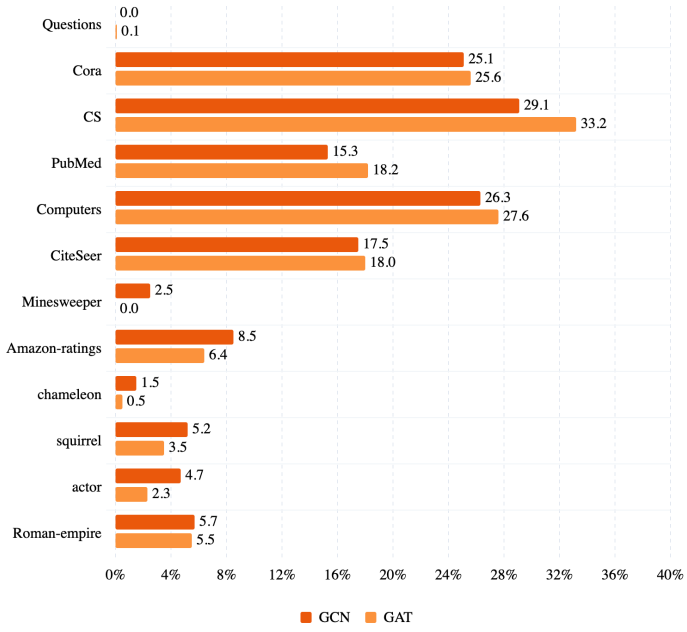


Figure 8: **Changed prediction under homophily reduction for GCN vs. GAT.** The plot reports the percentage of nodes whose predictions changed after the full homophily-reduction perturbation is applied.

G LLM USAGE

LLM was used to refine writing https://chatgpt.com/share/69238e52-c168-8013-920c-060dbad58155