PAIRSHIFT: A CONTROLLED BENCHMARK FOR GRAPH FOUNDATION MODELS UNDER DENSITY SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating the structural robustness of graph foundation models (GFMs) often yields a paradox: poorly performing models can appear deceptively robust, whereas highly capable models often prove brittle. Existing cross-dataset evaluations fail to isolate this failure mode, as topology, features, labels, and sampling confound one another. In this paper, we introduce PAIRSHIFT, a benchmark of *paired density views*: nested edge sets that share nodes, features, labels, and edge randomness. Its context-by-query response surfaces and an exact decomposition distinguish ordinary density effects from a matched-density interaction. Across six GFMs, G2T-FM and GraphPFN incur 16–17 percentage points of mean transfer failure on the common grid, rising to 19 points for G2T-FM on an unseen universe; the other four systems show much smaller interactions, sometimes alongside weak accuracy or unstable predictions. Fixed-checkpoint interventions then show that GraphPFN's graph-convolution adapters increase mean transfer failure by 11–44 points across controlled synthetic settings, with effects that depend on shift direction and adapter depth. Motivated by this pathway, we construct a class-compatibility neighbor residual with no learned parameters. It improves G2T-FM and SAMGPT across two evaluated universes (2–3 and 1–2 points, respectively) and GOBLIN by 4.4 points in a post-hoc extension, but harms models when additional neighbor evidence is redundant or unreliable. A local error-alignment analysis predicts the direction of its cross-entropy change in all 24 decisive tested conditions. PAIRSHIFT thus connects controlled structural evaluation to a causal failure pathway and a bounded, mechanism-inspired correction.

1 INTRODUCTION

Graph foundation models (GFMs) promise to transfer across diverse graph domains through in-context inference, prompting, or lightweight adaptation (Liu et al., 2023; Mao et al., 2024; Xia et al., 2024). A central premise of these approaches is structural transferability: can a model that learns from or adapts to context graphs generalize when the target query graph exhibits a markedly different edge density? In practice, evaluating this transfer reveals a deceptive paradox. A weak GFM that largely ignores topological signals can appear deceptively robust to density shifts, simply because it has little performance left to lose. Conversely, a capable model that relies heavily on graph structure may suffer a catastrophic performance drop whenever query density deviates from context density. Consequently, a small transfer gap is not, by itself, evidence of reliable structural transfer. Because graph adjacency dictates both the message-passing pathway and the computational flow (Gilmer et al., 2017; Xu et al., 2019), changing edge density fundamentally alters the model's internal computation—turning context-query structural mismatch into a critical yet obscured vulnerability.

Existing cross-dataset evaluations cannot isolate this problem: node identities, feature semantics, labels, topology, scale, and sampling often change together (Gui et al., 2022; Chen et al., 2024). The same structural change can also enter different GFMs through a compressed neighborhood table, a prompt, a learned graph, or repeated graph-convolution layers. What we need to measure is not merely whether sparse and dense graphs have different accuracies. It is whether a model pays an additional penalty when context and query densities disagree, compared with a density-matched context evaluated on the *same query graph*.

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.

We introduce PAIRSHIFT, a benchmark built around *paired density views* (Section 3). It fixes nodes, features, labels, and potential-edge randomness within each latent graph while varying only the edge-inclusion threshold, yielding nested views without independent edge resampling. Crossing disjoint context and query instances over a common density grid produces response surfaces with a matched-context control for every query density. Read alongside matched accuracy and prediction stability, an exact two-way decomposition separates ordinary density effects from the matched-density interaction.

Across six GFMs, PAIRSHIFT reveals behaviors that a scalar transfer score obscures (Section 3.4). The capable G2T-FM and GraphPFN incur 16–17 percentage points of mean transfer failure on the common grid, rising to 19 points for G2T-FM on an unseen universe; GOBLIN remains accurate with a small, smooth penalty; and AnyGraph's near-zero mean masks sign-changing responses, weak accuracy, and unstable predictions. A small average gap can therefore indicate genuine stability, limited capability, or cancellation across shift directions.

To move from system-level comparison to mechanism, we intervene within a fixed GraphPFN checkpoint (Section 4). An a priori *structural-support score* predicts cell-level transfer failure, while layer-wise interventions identify the repeated raw-adjacency adapters as a causal carrier: disabling them removes the effect, and reactivation increases its mean by 11–44 points under topologically informative conditions. This adapter contribution depends on shift direction and depth, while thinned real-graph tests show that its magnitude also depends on label–topology alignment and edge structure.

Motivated by this pathway, we construct a class-compatibility neighbor residual with no learned parameters (Section 4.2). Its local cross-entropy analysis predicts improvement when alignment with the base model's errors exceeds the curvature cost and matches the direction of change in all 24 decisive tested conditions. The residual improves G2T-FM and SAMGPT, as well as GOBLIN in a post-hoc extension, but becomes neutral or harmful for GraphPFN, MDGFM, and AnyGraph when related neighbor evidence is already absorbed or unreliable. It is therefore a bounded, mechanism-inspired correction rather than a universal repair of the matched-density interaction.

In conclusion, our contributions are as follows:

- **A controlled structural benchmark.** PAIRSHIFT uses nested density views and response surfaces to separate capability and stability from a matched-density interaction.
- **A causal failure pathway.** Fixed-checkpoint interventions identify GraphPFN's graph adapters as a carrier of this interaction and locate the boundary of that account.
- **A bounded structural correction.** A parameter-free neighbor residual and its error-alignment analysis explain when omitted structural evidence helps or harms.

2 RELATED WORK

Graph foundation models and cross-graph transfer. Graph foundation models (GFMs) seek predictors that transfer across graphs, tasks, and feature spaces (Liu et al., 2023; Mao et al., 2024). Graph-native approaches construct reusable interfaces through prompt graphs or nodes of interest (Huang et al., 2023; Liu et al., 2024), computation and task trees (Wang et al., 2024; 2025c), topology alignment and expert coordination (Yu et al., 2025; Wang et al., 2025a; Zhao et al., 2024; Yuan et al., 2025), or shared tokenizers and expert banks (Xia et al., 2024; Xia & Huang, 2026). A complementary line uses language as the common representation (He et al., 2025; Kong et al., 2025; Lin et al., 2024; Chen et al., 2024). These works establish broad transfer by evaluating on different datasets and tasks, where topology typically changes together with features, labels, semantics, scale, and sampling.

Our benchmark asks an orthogonal question: how does a fixed pretrained model respond when the latent task is held constant and only nested edge realizations change? We study six text-free GFMs that accept arbitrary numerical features—G2T-FM, GraphPFN, SAMGPT, MDGFM, GOBLIN, and AnyGraph (Eremeev et al., 2025; 2026; Yu et al., 2025; Wang et al., 2025a; Gutteridge et al., 2026; Xia & Huang, 2026)—and treat their different topology interfaces (Table 1) as an axis for mechanism analysis.

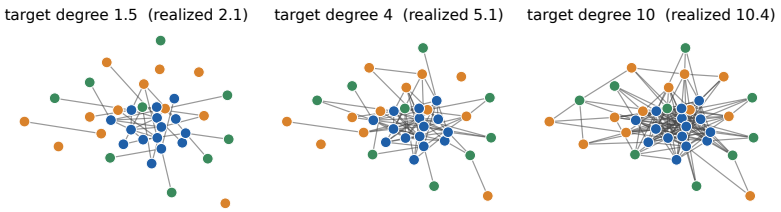

Figure 1: Three density views of the same underlying graph task. Nodes (colored by community), labels, features, degree propensities, and the uniform U_{ij} of every potential edge are shared; only the threshold $p_{ij}(d)$ changes. Lower-density edge sets are therefore subsets of higher-density ones.

Controlled structural shift. Graph OOD benchmarks distinguish covariate from concept shift across environments (Gui et al., 2022), while architecture search under distribution shift studies which GNN designs transfer best (Qin et al., 2022). Theory and controlled experiments further show that transfer depends on the local structures seen during training and on the relation between graph samples (Yehudai et al., 2021; Ruiz et al., 2020). Separately, heterophily work shows that edge homophily alone does not determine whether aggregation helps: the class-compatibility pattern matters (Zhu et al., 2020; Ma et al., 2022). These lines explain important source–target failures, but they do not provide the paired context-by-query design needed to isolate a matched-density interaction.

From structural diagnosis to correction. Methods for graph shift commonly learn a more invariant predictor or adapt the model to the target graph; examples include causal interventions during training (Wu et al., 2024) and self-supervised test-time training on each test graph (Wang et al., 2025b). A different tradition leaves the base predictor intact and propagates labels or predictions over the observed graph (Zhou et al., 2003); Correct & Smooth, for example, corrects labeled residuals before smoothing predictions (Huang et al., 2021). Unlike learned adaptation or generic smoothing, our parameter-free one-step residual serves as a bounded probe of a topology pathway identified through fixed-checkpoint intervention.

3 BENCHMARK

PAIRSHIFT constructs multiple density views of the same synthetic node-classification task, then asks a controlled question: if a GFM is conditioned on context graphs at density d_c, how does it perform on query graphs at density d_q when the underlying prediction problem is otherwise unchanged? Paired views vary only the observed edge set; a context-by-query grid reveals the full response surface; and a density-matched control turns every shifted cell into a within-query contrast. We report matched capability, transfer failure, and prediction stability together, so a model cannot appear robust merely because it performs poorly or changes its predictions unpredictably.

3.1 PAIRED GRAPH CONSTRUCTION

PAIRSHIFT separates two levels of construction. A *universe* is a synthetic environment shared by many graphs: it specifies a pool of communities, their relative edge and degree propensities, and the class-conditional feature distributions. From this environment we sample multiple underlying node-classification tasks. A *latent instance* z is one such task before an observation density is chosen. It fixes the nodes, participating communities, labels $\boldsymbol{y}_z$, features $\boldsymbol{X}_z$, node degree propensities, and an independent uniform $U_{ij,z}$ for every potential edge.

Rendering instance z at density d changes only which of those potential edges are observed:

$$\mathcal{G}_z(d) = \big(\boldsymbol{A}_z(d), \boldsymbol{X}_z, \boldsymbol{y}_z\big), \qquad A_{ij,z}(d) = \mathbf{1}\{U_{ij,z} \le p_{ij,z}(d)\}, \tag{1}$$

where the density-specific edge probability $p_{ij,z}(d)$ grows with target average degree. Because every view reuses the same $U_{ij,z}$, $d_1 < d_2$ implies $E_z(d_1) \subseteq E_z(d_2)$ (Figure 1). A denser view therefore reveals additional edges of the same task rather than resampling a new graph.

Table 1: Topology paths in the benchmark implementations. *Context use* describes how labeled context enters target-time prediction.

Model	Context use	Topology path
G2T-FM	in-context inference	graph-derived features before LimiX
GraphPFN	in-context inference	raw-adjacency residual adapters after each block
SAMGPT	prompt optimization	prompt-modulated GCN layers
MDGFM	prompt/topology optimization	GCN on a raw–learned topology mixture
GOBLIN	operator search + analytic experts	context-selected $S(A)$ operators
AnyGraph	expert routing + class nodes	adjacency SVD + topology encoder

We sample many latent instances from each universe and split them by identity into context, validation, and test sets. Context and query never reuse an instance; the pairing is across density views of each instance, not across the context–query split. Four structural strata vary homophily and degree-tail shape, and every instance is rendered at five densities. Appendix A gives the complete parameters, sample sizes, validity gate, and invariant audit. This construction enables Eq. 2 to change context density while holding the query task and query view fixed.

3.2 Context-by-query response surfaces

With these paired views in hand, we cross the density of the context graphs with that of the held-out query graphs. For model m, structural stratum s, context density d_c, and query density d_q, let $a_m(s, d_c, d_q)$ denote test accuracy averaged over model or adaptation seeds. For example, $a_m(s, 2, 32)$ evaluates degree-32 query views after conditioning on degree-2 context views, whereas $a_m(s, 32, 32)$ uses the matched context on the same query views. Crossing all five context and query densities produces a response surface whose diagonal is matched and whose off-diagonal cells are shifted. For each shifted cell, transfer failure is

$$\mathcal{F}_m(s, d_c \to d_q) = a_m(s, d_q, d_q) - a_m(s, d_c, d_q). \tag{2}$$

A positive value means that replacing the matched context with one at density d_c reduces accuracy. Because both terms use the same query graph and test nodes, Eq. 2 does not confuse context mismatch with a change in query difficulty or an independent edge sample.

We call Eq. 2 the cell-level *transfer failure* and its average over off-diagonal cells the *mean transfer failure*. The benchmark output is the full response surface, not this average alone. We report matched and off-diagonal accuracy to retain capability and the signed within-query contrast to retain shift direction. As a complementary stability measure, we also change query density under a common context and record paired prediction flips; unlike transfer failure, this measure perturbs the query view. The benchmark results below use an exact decomposition to distinguish ordinary context or query density effects from an additional matched-density interaction.

3.3 Models and evaluation protocol

We evaluate six GFMs whose interfaces expose topology in different ways (Table 1). Each system uses its pinned public implementation and documented multi-graph interface, with disjoint context and query latent identities (Appendix A.2). Sampling units and intervals are defined over latent graphs or other declared clusters, never over correlated nodes. Confirmatory claims require every check in a pre-specified set to pass; failed sets and studies added after the freeze are reported with their status (Appendix A.4).

3.4 Benchmark results: capability, failure, and stability

The definitions above map directly to the reported summaries. Figure 2 plots $\mathcal{F}_m(s, d_c \to d_q)$ averaged over structural strata and seeds. Table 2 reports the corresponding mean diagonal accuracy

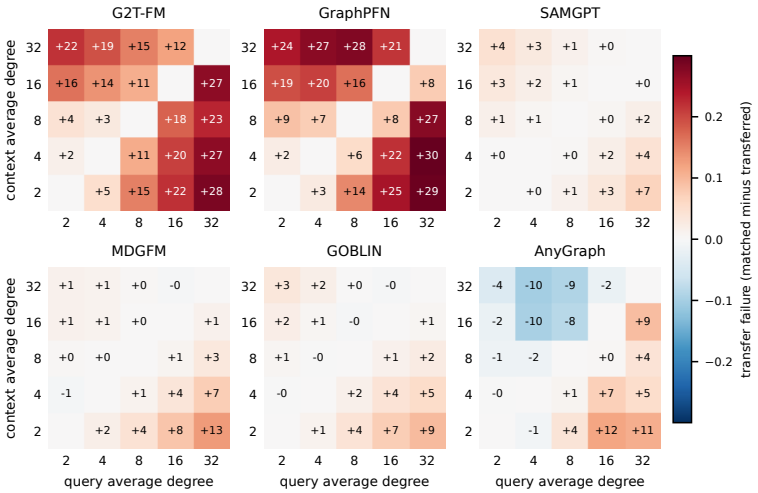

Figure 2: Transfer-failure surfaces (Eq. 2, accuracy points; rows context density, columns query density; four strata and three seeds averaged). The same paired evaluation gives a steep, direction-dependent surface for G2T-FM and GraphPFN, a nearly flat one for SAMGPT, MDGFM, and GOBLIN, and a sign-changing one for AnyGraph.

(ID), mean off-diagonal accuracy (OOD), mean off-diagonal transfer failure, and the prediction-flip rate between the two extreme query views under a common context.

Every GFM sees the same latent tasks and paired grid, yet the surfaces in Figure 2 differ in shape, not just in scale (Table 2). G2T-FM and GraphPFN combine high matched accuracy with the largest drop, up to 30 points sparse-to-dense; GOBLIN has the highest capability with a small, smooth drop; SAMGPT and MDGFM lose almost nothing from a weaker starting point; and AnyGraph's near-zero mean comes from negative and positive cells cancelling around a starting point close to chance, while its paired predictions stay unstable. The GCN and GraphSAGE rows are task-specific supervised references, not compute- or supervision-matched competitors: each is trained from scratch at every context density using all node labels in the 60 context graphs. Their competitive or higher matched accuracy but much smaller mean failure shows that the G2T-FM and GraphPFN surfaces are not merely scaled-up GNN surfaces.

Table 2: Common-grid response map. ID/OOD are mean diagonal/off-diagonal accuracy; Failure is mean transfer failure from Eq. 2. Extreme flip is complementary: it compares degree-2 and degree-32 query predictions under a common context.

Model	ID acc.	OOD acc.	Failure	Extreme flip
G2T-FM	0.626	0.470	0.157	0.55
GraphPFN	0.600	0.427	0.173	0.68
SAMGPT	0.489	0.471	0.019	0.61
MDGFM	0.431	0.407	0.024	0.58
GOBLIN	0.654	0.632	0.022	0.31
AnyGraph (stochastic projection)	0.293	0.290	0.002	0.62
AnyGraph (deterministic projection)	0.341	0.344	−0.003	0.70
GCN (supervised reference)	0.651	0.626	0.026	0.52
GraphSAGE (supervised reference)	0.730	0.674	0.055	0.37

The two AnyGraph rows separate its original stochastic projection from a deterministic audit; the former is retained because Section 5 uses its cached logits (Appendix A.2).

Exact response-surface decomposition. A scalar failure average can conflate a matched-density interaction with ordinary context or query effects. We therefore decompose each surface as $a(d_c, d_q) = \mu + \alpha_{d_c} + \beta_{d_q} + \gamma_{d_c d_q}$ under zero-sum constraints. Mean transfer failure is exactly

$\overline{\gamma}_{\mathrm{diag}} - \overline{\gamma}_{\mathrm{off}}$ and vanishes for an additive surface (Appendix B). The interaction is 0.16–0.17 for G2T-FM and GraphPFN, whereas it is -0.003 to 0.024 and main effects explain 78–89% of cell variance for GOBLIN, MDGFM, and deterministic AnyGraph. All six interactions replicate on an unseen universe (Appendix C).

3.5 Structural support beyond nominal density

The response surfaces leave two questions: which realized structural changes predict failure, and which model exposes a controllable pathway through which that failure enters computation? Density labels alone do not describe how a particular graph changes. For each directed density pair, we therefore form a 19-dimensional *structural-support vector* of realized changes in connectivity, isolation, degree distribution, label mixing, feature aggregation, and global structure. The label-mixing coordinate uses context labels, so this is a diagnostic rather than a query-time rule. For each architecture, a ridge predictor from this vector to Eq. 2 and a density-only baseline were fitted on one development universe and frozen before evaluation on a new universe.

The result is model-specific (Table 3). For GraphPFN, structural support reaches $R^2 = 0.75$ and $\rho = 0.92$, improving R^2 over density alone by 0.48 $[0.36, 0.62]$; without refitting, the gain remains 0.21–0.41 on four further universes. MDGFM is predictable but gains only 0.06 beyond density; G2T-FM gains 0.07 with an interval crossing zero; and SAMGPT preserves rank but not calibrated scale. GOBLIN passes the same prediction gate, whereas deterministic AnyGraph improves R^2 but not rank over density alone. Because the predictors are architecture-specific, they do not define a universal cross-model score. We intervene on GraphPFN because it combines strong mean transfer failure, prospective structural predictability, and a separable topology path.

Table 3: Frozen architecture-specific structural-support predictors evaluated on held-out universes. R^2 measures absolute predictive fit, ρ rank ordering, and ΔR^2 the gain over density alone.

Model	Composite R^2	Spearman ρ	ΔR^2 over density
GraphPFN	0.75	0.92	$+0.48$ [0.36, 0.62]
G2T-FM	0.34	0.59	$+0.07$ [-0.09, 0.24]
SAMGPT	-2.07	0.74	-0.62 [-1.04, -0.34]
MDGFM	0.74	0.85	$+0.06$ [0.02, 0.09]
GOBLIN	0.77	0.92	$+0.06$ [0.02, 0.11]
AnyGraph (det.)	0.37	0.74	$+0.13$ [0.06, 0.19]

4 From GraphPFN's Failure Pathway to a Residual Correction

Section 3.5 makes GraphPFN the clearest mechanism case: it is accurate when densities match, fails strongly when they do not, and exposes a topology path that can be disabled without retraining. Structural support tells us what predicts failure; this access lets us ask how failure enters the computation and extract a simple correction.

4.1 Why GraphPFN is sensitive to density mismatch

GraphPFN alternates a context-conditioned transformer with a graph adapter,

$$\boldsymbol{H}_{\ell+1} = T_\ell(\boldsymbol{H}_\ell; \mathcal{C}) + G_\ell(\boldsymbol{H}_\ell, \boldsymbol{A}_q). \tag{3}$$

Context density enters T_ℓ only through adapter-updated context states: graph convolution changes those states, global attention transfers the change to the query, and later adapters propagate the resulting states over the fixed query adjacency $\boldsymbol{A}_q$. With all adapters disabled, paired context views have identical transformer-visible inputs. Repetition can therefore amplify context–query mismatch, which PairShift isolates while holding the query view fixed.

We test this explanation inside one frozen checkpoint, retaining 0, 6, or all 12 adapters. At depth zero, predictions are exactly invariant to context density. Restoring all adapters increases mean transfer failure by 0.11 $[0.09, 0.14]$: matched accuracy falls by 0.04, but cross-density accuracy falls by 0.15.

This depth-12 minus depth-0 contrast is the *adapter contribution*; it shows that the adapters magnify mismatch rather than simply lowering accuracy.

This depth contrast remains 0.11–0.44 across new universes and generator settings. A factorial intervention points to graph convolution rather than the adapter MLP, although its full gate fails. On real graphs the contrast is small with original labels; planting the generator's label process on the same edges restores 0.05–0.17 on three of four graphs. The pathway matters when topology carries the label signal, not whenever a graph is present (Appendix C).

4.2 Extracting the simplest useful operation

Removing the adapters also discards useful topology; the causal result does not imply that aggregation is intrinsically harmful. We therefore test whether one output-space message can add reliable neighbor evidence without recreating the repeated, context-conditioned pathway. The residual collects neighbor evidence in label space, matches it to the model's logit scale, and adds it once at the output. It is a mechanism-inspired probe rather than an inverse of GraphPFN's hidden-state computation, and applies to any model producing query probabilities.

Let $\boldsymbol{P}$ be those probabilities, $\boldsymbol{A}_q$ the adjacency with degree matrix $\boldsymbol{D}$, and $\boldsymbol{T}$ a smoothed class-transition matrix estimated only from labeled context edges. From hard query predictions $\widehat{\boldsymbol{Y}} = \mathrm{onehot}(\arg\max_k \boldsymbol{P})$, we compute

$$\boldsymbol{E} = \boldsymbol{D}^{-1}\boldsymbol{A}_q\widehat{\boldsymbol{Y}}\log(\boldsymbol{T})^\top, \tag{4}$$

$$\widetilde{\boldsymbol{e}}_i = \mathrm{center}(\boldsymbol{e}_i)\,\frac{\mathrm{sd}(\mathrm{center}(\log \boldsymbol{p}_i))}{\max\{\mathrm{sd}(\mathrm{center}(\boldsymbol{e}_i)), \epsilon\}}, \tag{5}$$

$$\boldsymbol{p}'_i = \mathrm{softmax}\big(\mathrm{center}(\log \boldsymbol{p}_i) + \tfrac{1}{2}\widetilde{\boldsymbol{e}}_i\big)\,. \tag{6}$$

In the first line, neighbors vote through their pseudo-labels and $\log(\boldsymbol{T})$ converts each vote into evidence for every candidate class; $\boldsymbol{D}^{-1}$ averages the votes so degree does not mechanically enlarge the update. The second line centers and matches this message to the node's base-logit scale. The last adds half of it and returns to probabilities. This fixed step uses no query labels, graph edits, learned weights, or model-specific tuning.

4.3 Prediction and tests

The residual should help only when its neighbor evidence is reliable and has not already been absorbed. For logits $\boldsymbol{Z}_m + \eta\boldsymbol{R}$, cross-entropy changes locally as

$$\mathcal{L}_m(\eta) - \mathcal{L}_m(0) = -\eta A_m + \tfrac{1}{2}\eta^2 V_m + O(\eta^3), \quad A_m = \mathbb{E}\langle \boldsymbol{Y} - \boldsymbol{P}, \boldsymbol{R}\rangle, \quad V_m \geq 0. \tag{7}$$

A_m asks whether the residual points toward the remaining error; V_m penalizes a finite step. Omitted reliable evidence should make A_m positive. If it is already absorbed, curvature can dominate; if pseudo-labels are wrong, A_m can be negative. Because this calculation uses held-out labels, it explains cross-entropy after evaluation rather than providing a deployment rule or accuracy guarantee.

Section 5 tests four consequences: cross-model accuracy gains and losses; reversal as GraphPFN's adapters are restored; contrasting behavior in GOBLIN and AnyGraph; and agreement with Eq. 7 and standard smoothers. Differences are paired on the same latent query graphs, isolating the residual from task variation. These tests assess a conditional algorithmic consequence of the pathway; they do not by themselves show that the residual reduces mean transfer failure. Full settings are in Appendices A and C.

5 When Does the Residual Help?

We answer the four questions above in order. Every difference compares corrected and uncorrected predictions on the same query instance. The outcome here is accuracy or cross-entropy change, not a direct change in mean transfer failure. Figure 3 summarizes the benchmark–mechanism–correction chain.

The hypothesis has two separate requirements: the neighbor message must be reliable, and the base model must not already contain equivalent evidence. Cross-model tests probe both at once; the GraphPFN depth intervention isolates the second.

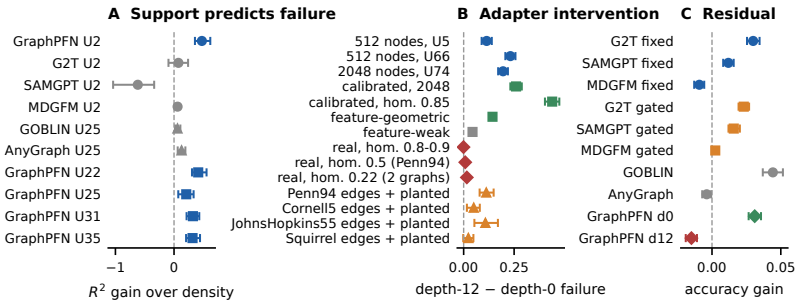

Figure 3: Evidence chain: structural support predicts failure; activating GraphPFN's adapters creates it; and the residual helps only when reliable neighbor evidence remains unused. Means and 95% cluster-bootstrap intervals.

5.1 Experiment 1: Cross-model evaluation

In the first prospective cross-model test, we apply the same correction, with no model-specific fitting, to G2T-FM, SAMGPT, and MDGFM. It raises accuracy by 3.0 points for G2T-FM and 1.2 for SAMGPT, but lowers MDGFM by 0.9 (Table 4). The negative result is informative: all three models receive the same query topology, yet only two produce neighbor votes aligned with their remaining errors. Merely having more edges is therefore not enough.

Table 4: Residual-minus-base accuracy with 95% latent-cluster intervals. All rows use paired query instances. † marks post-hoc cross-architecture extensions.

Two-universe evaluation			Model and pathway contrasts	
Model	Frozen, universe 1	Gated, universe 2	Model or condition	Fixed residual
G2T-FM	$+0.030$ [0.025, 0.035]	$+0.023$ [0.019, 0.027]	GOBLIN†	$+0.044$ [0.037, 0.052]
SAMGPT	$+0.012$ [0.008, 0.016]	$+0.016$ [0.012, 0.020]	AnyGraph†	-0.004 [-0.007, -0.001]
MDGFM	-0.009 [-0.013, -0.006]	$+0.002$ [0.000, 0.005]	GraphPFN, depth 0	$+0.031$ [0.027, 0.036]
Pooled	$+0.011$ [0.008, 0.014]	$+0.014$ [0.012, 0.016]	GraphPFN, depth 12	-0.015 [-0.019, -0.011]

The directions explain this split. At homophily 0.5, sparse-to-dense gains reach $+6.8$ and $+3.4$ points for G2T-FM and SAMGPT; at homophily 0.2, less reliable votes leave SAMGPT neutral and cost MDGFM 2.6 points. On a second outcome-untouched universe, a fixed headroom gate compares context labeled-edge agreement with query pseudo-label agreement. A large gap means that the query predictions are less locally consistent than the context labels suggest they could be, so the residual is applied; otherwise it abstains. This retains gains for G2T-FM ($+2.3$) and SAMGPT ($+1.6$) and leaves MDGFM near zero. Because improvement over the ungated update is not established for MDGFM, the gate reduces risk but is not a universal selector.

5.2 Experiment 2: Controlled evidence absorption

To remove cross-model confounds, we change only one pathway inside GraphPFN, holding its checkpoint and query graphs fixed. With 0 adapters, the residual gains 3.1, 2.5, and 5.9 points across three universes; with all 12, these become -1.5, -0.5, and -0.4. Every paired depth contrast excludes zero. The same evidence therefore helps before graph propagation and becomes redundant or harmful afterward.

At depth six, accuracy still improves by 1.6–4.9 points although NLL already worsens by 0.13–0.36. The residual is beginning to duplicate evidence: it can still flip some incorrect argmax decisions, but it also makes other predictions overconfident. Calibration therefore fails before top-1 accuracy. An adjacency-blind MLP does not benefit because its pseudo-labels are weak. Depth-zero gains are also 4.9–10.1 points sparse-to-dense but only 0.1–1.7 in reverse. Missing topology is insufficient; the query must supply reliable new evidence. Full dose and gating results are in Appendix C.

5.3 Experiment 3: Does topology use alone predict benefit?

In these post-hoc extensions, GOBLIN and AnyGraph both process topology, yet respond oppositely. For GOBLIN, accuracy rises by 4.4 points and NLL falls by 0.052; the $+8.8$-point sparse-to-dense gain and near-zero reverse effect show that accurate pseudo-labels leave useful compatibility evidence precisely when dense query neighborhoods provide it. For AnyGraph, accuracy falls by 0.4 points and NLL rises by 0.76. Sparse-to-dense accuracy falls by 1.4 points; reverse transfer gains 0.6 points but still worsens NLL. Its neighbors often agree because they are jointly wrong, so local consistency is not label reliability and the gate abstains throughout. Native topology use therefore predicts neither outcome: the particular output evidence must be reliable and missing.

5.4 Experiment 4: Error alignment and smoother controls

Equation 7 predicts the NLL-change sign in all 24 decisive conditions, meaning those whose observed intervals exclude zero ($\rho = 0.985$). Alignment also decreases with GraphPFN depth, exactly as expected when its adapters progressively absorb the same evidence. It predicts accuracy less well ($\rho = 0.50$): G2T-FM, for example, gains 3.0 accuracy points while NLL worsens by 0.12. The equation therefore explains the local calibration response, not whether an argmax will change correctly or whether the residual should be deployed.

For G2T-FM, the residual's pooled gain is 2.7 points, versus 1.7 for label propagation and 1.1–1.2 for Correct & Smooth; for SAMGPT, plain propagation is competitive (1.8 versus 1.4). A second residual step helps wherever one step helps, but every propagation variant harms MDGFM. These controls place the method within the smoothing family: its contribution is the context-derived compatibility message and an interpretable boundary, not universal superiority.

6 Discussion and Limitations

Paired surfaces show why a small scalar gap need not imply robustness; within GraphPFN, fixed-checkpoint interventions identify adapters as one causal pathway when topology is predictive. The residual asks the narrower question of whether reliable neighbor evidence remains unused, not whether it universally repairs matched-density failure.

The strongest matched-density and adapter-path effects occur in controlled synthetic universes. Extensions to thinned real graphs show only a small trace and indicate that the pathway depends on both whether task labels are carried by the observed topology and whether the model's adapters can use the particular edge structure (Appendix D). PairShift therefore establishes a failure mode that can be isolated and studied, not its prevalence or magnitude across naturally occurring graph tasks. The synthetic universes also share a latent seed across resampled structural and feature parameters, and the largest-scale replications use a calibrated generator variant (Appendix A).

The fixed-checkpoint interventions show that GraphPFN's graph adapters carry the effect, but do not identify the exact internal mediator; the tested norm and orientation accounts were insufficient. Cross-model contrasts remain descriptive because architecture, pretraining, objective, and adaptation vary together, and some mechanism studies use one model seed. Finally, the residual is not a universal deployment rule: the compound safety gate failed, while error alignment uses held-out labels only for retrospective analysis and predicts cross-entropy rather than accuracy.

7 Conclusion

In this paper, we introduced PairShift, a benchmark of paired density views whose response surfaces separate ordinary density effects from a matched-density interaction across six GFMs. Fixed-checkpoint interventions identify GraphPFN's graph-convolution adapters as one causal carrier, while a parameter-free residual yields replicated gains when reliable neighbor evidence remains unused. Structural robustness should therefore be evaluated by how a model transforms topology, not by an undifferentiated cross-dataset transfer gap.

AI USE STATEMENT

Generative AI tools assisted with conceptual framing, theoretical analysis, experimental design and orchestration, code development, result interpretation and checking, and manuscript drafting and editing. All reported measurements come from stored program outputs under versioned protocols. The authors reviewed all AI-assisted work, verified numerical and literature claims, and take responsibility for the final content.

ETHICS STATEMENT

This work is methodological and does not deploy predictions or make consequential decisions. Experiments use synthetic graphs and established public benchmark datasets; no new human-subject data are collected. We use datasets and model implementations under their stated research terms.

REPRODUCIBILITY STATEMENT

Section 3 defines the benchmark estimands and evaluation protocol. Appendix A documents the generator parameters, model versions, sampling units, and statistical procedures; Appendices B–D provide derivations, supporting evidence, and real-graph controls. All reported results were produced from versioned configurations and frozen analysis protocols. Code, configurations, and experiment manifests will be released publicly upon acceptance.

REFERENCES

Zhikai Chen, Haitao Mao, Jingzhe Liu, Yu Song, Bingheng Li, Wei Jin, Bahare Fatemi, Anton Tsitsulin, Bryan Perozzi, Hui Liu, and Jiliang Tang. Text-space graph foundation models: Comprehensive benchmarks and new insights. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0239.

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

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

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–1272, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Benjamin Gutteridge, Michael Bronstein, and Xiaowen Dong. Can graph foundation models generalize over architecture? In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026.

Yufei He, Yuan Sui, Xiaoxin He, and Bryan Hooi. UniGraph: Learning a unified cross-domain foundation model for text-attributed graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 448–459, 2025.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy Liang, and Jure Leskovec. PRODIGY: Enabling in-context learning over graphs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. GOFA: A generative one-for-all model for joint graph language modeling. In *International Conference on Learning Representations*, 2025.

Louis Van Langendonck, Guillermo Bernardez, Nina Miolane, and Pere Barlet-Ros. Graphuniverse: Synthetic graph generation for evaluating inductive generalization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tianqianjin Lin, Pengwei Yan, Kaisong Song, Zhuoren Jiang, Yangyang Kang, Jun Lin, Weikang Yuan, Junjie Cao, Changlong Sun, and Xiaozhong Liu. LangGFM: A large language model alone can be a powerful graph foundation model. *arXiv preprint arXiv:2410.14961*, 2024.

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

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. Towards graph foundation models: A survey and beyond. *arXiv preprint arXiv:2310.11829*, 2023.

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

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34670–34692, 2024.

Yijian Qin, Xin Wang, Ziwei Zhang, Pengtao Xie, and Wenwu Zhu. Graph neural architecture search under distribution shifts. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 18083–18095, 2022.

Luana Ruiz, Luiz Chamon, and Alejandro Ribeiro. Graphon neural networks and the transferability of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1702–1712, 2020.

Shuo Wang, Bokui Wang, Zhixiang Shen, Boyan Deng, and Zhao Kang. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 64806–64821, 2025a.

Yiqi Wang, Chaozhuo Li, Jianan Zhao, and Rui Li. Test-time training for graph neural networks. In *International Conference on Database Systems for Advanced Applications*, pp. 349–364. Springer, 2025b.

Zehong Wang, Zheyuan Zhang, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. GFT: Graph foundation model with transferable tree vocabulary. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3412.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Towards graph foundation models: Learning generalities across graphs via task-trees. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 65518–65555, 2025c.

Qitian Wu, Fan Nie, Chenxiao Yang, Tianyi Bao, and Junchi Yan. Graph out-of-distribution generalization via causal intervention. In *Proceedings of the ACM Web Conference*, pp. 850–860, 2024.

Lianghao Xia and Chao Huang. AnyGraph: Graph foundation model in the wild. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 882–896, 2026.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11975–11986, 2021.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. SAMGPT: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM Web Conference*, 2025.

Haonan Yuan, Qingyun Sun, Junhua Shi, Xingcheng Fu, Bryan Hooi, Jianxin Li, and Philip S. Yu. How much can transfer? BRIDGE: Bounded multi-domain graph foundation model with generalization guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 73604–73644, 2025.

Haihong Zhao, Aochuan Chen, Xiangguo Sun, Hong Cheng, and Jia Li. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4443–4454, 2024.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pp. 321–328, 2003.

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 Experimental and Statistical Details

A.1 Paired synthetic universes

Our paired generator extends GraphUniverse (Langendonck et al., 2026) to render each latent graph at multiple nested densities.[1]

Each universe defines $K = 10$ communities and shared edge, degree, and feature parameters. Every latent instance samples five participating communities and is rendered at several densities. Its views share node order, labels and community assignments, node features, degree propensities, and one uniform random variable for every potential edge; changing density only changes the threshold applied to those uniforms. Lower-density edge sets are therefore exact subsets of higher-density ones. The primary grid crosses target homophily $\{0.2, 0.5\}$, power-law exponents $\{2.5, 4.2\}$, and target average degrees $\{2, 4, 8, 16, 32\}$. Each stratum has 100 graphs of 512 nodes with 32-dimensional class-conditional features, split by latent identity into $60/20/20$ context, validation, and test sets using seed 17. Model or adaptation seeds are $0, 1, 2$ unless a stated intervention fixes one seed.

Before model inference, every cell must pass outcome-free checks for exact edge nesting and shared attributes, edge-probability clipping at most 0.05, relative average-degree error at most 0.25, absolute homophily error at most 0.15, and checksum and coverage consistency. Candidate universes are examined in a predeclared order; failure advances to the next candidate without reading model outcomes.

Across universe seeds, the community propensity matrix, degree propensities, and feature centers are resampled, but a common latent seed preserves community assignments, labels, and edge uniforms for a given latent ID. These are thus new parameter draws from the same family rather than fully independent latent populations. Replications above $1{,}000$ nodes use a calibrated variant that normalizes community-block weights exactly; the original 512-node family is unchanged.

A.2 Models and interventions

All GFM runs use pinned public sources: GraphPFN commit `3b9b115` with checkpoint 1.3, G2T-FM `e279c0f`, SAMGPT `5ca1`, MDGFM `495bfc`, GOBLIN `54c7c61`, and AnyGraph `c2c5bfe`. Versioned configurations record full checkpoint and parameter digests.

[1] `https://github.com/LouisVanLangendonck/GraphUniverse`

For GraphPFN, depth zero bypasses all graph adapters while retaining the transformer and is exactly invariant to adjacency under paired inputs; depths 6 and 12 activate the first six or all adapters. MLP-only and graph-convolution-only patterns isolate the two adapter branches without retraining. The supervised references in Table 2 are two-layer GCN and GraphSAGE models with hidden width 64 and dropout 0.2, trained with Adam and validation macro-F1 early stopping on all labels from the 60 context graphs; results average three seeds.

For the residual in Eq. 6, the context transition matrix uses additive smoothing of 1 before taking logarithms. Query pseudo-labels are frozen from the base argmax; degree normalization includes isolated-node safeguards; centered evidence is matched to the nodewise standard deviation of the base log probabilities. The strength is fixed at 0.5 and the update is applied once unless explicitly labeled as a two-step comparison.

A.3 Model selection

The original frozen model set comprised G2T-FM, GraphPFN, SAMGPT, and MDGFM; GOBLIN and AnyGraph were integrated later to broaden architectural coverage. All six benchmark evaluations use the same complete density grid, latent instances, estimands, seed counts, and uncertainty procedure, and all cells are reported under the same inclusion rules. The selection chronology therefore does not alter their benchmark protocol or comparability. Hypothesis-specific residual studies retain the prospective or post-hoc status defined by their own protocols.

A.4 Sampling units and uncertainty

All reported intervals use $10{,}000$ percentile bootstrap replicates over the declared independent unit, never over nodes. Response-surface summaries resample homophily–tail–seed clusters. The prospective support study resamples the 20 directed density pairs within each of four fixed strata. Adapter and residual studies first average model seeds within each held-out latent graph and then resample latent graphs; real-graph studies analogously resample query blocks. Paired contrasts are formed within the same latent graph or block before resampling.

Confirmatory language is used only when every component of a prespecified gate passes. The original support-prediction, adapter-removal, and first prospective residual gates passed. The adapter factorial and the five-part headroom-gate confirmation remain failed even where individual contrasts are positive. Later universe, scale, real-graph, smoother, and alignment studies are labeled post-freeze or descriptive in the supporting tables below.

B Analytical Derivations

B.1 Exact response-surface decomposition

Let m be the number of density levels and write $a_{cq} = \mu + \alpha_c + \beta_q + \gamma_{cq}$ under the standard zero-sum constraints. Averaging Eq. 2 over all ordered off-diagonal pairs gives

$$\overline{F} = \frac{1}{m(m-1)} \sum_q \sum_{c \neq q} (a_{qq} - a_{cq}) \tag{8}$$

$$= \frac{1}{m} \sum_q \gamma_{qq} - \frac{1}{m(m-1)} \sum_q \sum_{c \neq q} \gamma_{cq}. \tag{9}$$

The intercept and query main effect cancel within each difference. The context contribution is proportional to $\sum_q \alpha_q = 0$, leaving exactly the mean diagonal interaction minus the mean off-diagonal interaction. Hence $\overline{F} = 0$ for every additive surface. For two densities a and b, the directional asymmetry is

$$F(a \to b) - F(b \to a) = 2(\alpha_b - \alpha_a) + \gamma_{bb} - \gamma_{ab} - \gamma_{aa} + \gamma_{ba},$$

which separates a context-density main effect from an interaction remainder.

B.2 Local error alignment

For one observation with one-hot label $\boldsymbol{y}$, logits $\boldsymbol{z}$, probabilities $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$, and a fixed residual direction $\boldsymbol{r}$, cross-entropy satisfies

$$\nabla_{\boldsymbol{z}} \ell = \boldsymbol{p} - \boldsymbol{y}, \qquad \nabla^2_{\boldsymbol{z}} \ell = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \succeq 0.$$

A Taylor expansion along $\boldsymbol{z} + \eta \boldsymbol{r}$ therefore gives

$$\ell(\boldsymbol{z} + \eta \boldsymbol{r}) - \ell(\boldsymbol{z}) = -\eta \langle \boldsymbol{y} - \boldsymbol{p}, \boldsymbol{r} \rangle \tag{10}$$

$$+ \frac{\eta^2}{2} \left(\sum_k p_k r_k^2 - \left(\sum_k p_k r_k \right)^2 \right) + O(\eta^3). \tag{11}$$

Averaging yields Eq. 7; its curvature term is non-negative because it is the variance of r_k under $k \sim \boldsymbol{p}$. The derivation concerns cross-entropy along the fixed deployed direction. It neither guarantees an accuracy improvement nor supplies a label-free selector, since evaluating A_m requires held-out labels.

C Additional Evidence for the Main Claims

This appendix retains only results needed to audit claims made in the main text. Values were checked against the frozen and post-freeze experiment reports; complete cell-level outputs and exploratory studies are retained in the versioned experiment repository and will be released with the code.

C.1 Benchmark and predictor replication

The matched-density interaction replicates on a second universe for all six GFMs (Table 5). Additive surface type also replicates except for G2T-FM, whose interaction remains large but whose second-universe surface acquires a context main effect.

Table 5: Response-surface replication. Entries are matched-cell interaction / additive R^2.

Model	Universe 2	Universe 22
GraphPFN	0.173/0.19	0.156/0.35
G2T-FM	0.157/0.35	0.188/0.61
SAMGPT	0.019/0.66	0.023/0.66
MDGFM	0.024/0.86	0.034/0.69
GOBLIN	0.022/0.89	0.025/0.85
AnyGraph	0.002/0.78	0.006/0.80

The two post-freeze extensions in Table 3 use the same feature definition but separately frozen architecture-specific predictors. GOBLIN passes all six prospective checks. Deterministic AnyGraph passes five: its R^2 gain and MAE reduction are positive, but its rank-correlation gain over density has interval $[-0.06, 0.11]$. The original GraphPFN predictor was fitted only on universe 0 and applied without refitting. Its advantage over density-only prediction is positive on all five tested universes (Table 6); universe 25 misses one rank-correlation component of the original six-part gate but retains positive R^2 and MAE improvements.

Table 6: Prospective GraphPFN support prediction across universes.

Universe	Composite R^2	Spearman ρ	Density-only R^2	ΔR^2 [95% CI]
2	0.752	0.921	0.277	+0.475 [0.357, 0.619]
22	0.788	0.901	0.375	+0.413 [0.306, 0.555]
25	0.477	0.770	0.268	+0.208 [0.072, 0.334]
31	0.512	0.848	0.190	+0.322 [0.214, 0.431]
35	0.770	0.894	0.454	+0.316 [0.206, 0.443]

C.2 ADAPTER-PATH ROBUSTNESS

Table 7 collects the settings behind the 0.11–0.44 range in Section 4. Depth zero is exactly adjacency-invariant in every row. The feature-weak family is reported separately because its depth-zero accuracy is 0.108, close to chance; its smaller positive contrast does not show useful structural transfer.

Table 7: GraphPFN depth-12 minus depth-0 transfer failure across controlled settings.

Setting	Nodes	Contrast [95% CI]
Frozen family, universe 5	512	$+0.114$ $[0.089, 0.139]$
Frozen family, universe 9	512	$+0.209$ $[0.187, 0.230]$
Frozen family, universe 66	512	$+0.232$ $[0.207, 0.257]$
Frozen family, universe 74	2,048	$+0.195$ $[0.171, 0.219]$
Calibrated family, universe 61	1,024	$+0.231$ $[0.209, 0.254]$
Calibrated family, universe 61	2,048	$+0.260$ $[0.231, 0.288]$
Feature-geometric family	512	$+0.143$ $[0.128, 0.158]$
Homophily 0.5, ratio 16	512	$+0.237$ $[0.203, 0.269]$
Homophily 0.85, ratio 16	512	$+0.439$ $[0.403, 0.474]$
Feature-weak family	512	$+0.044$ $[0.031, 0.057]$

A six-pattern factorial localizes topology sensitivity to the convolution branch: MLP-only adapters are adjacency-invariant, whereas convolution-only adapters are sensitive but inaccurate. Late-adapter contributions and early–late interactions reverse with shift direction, so no block has a fixed context-independent penalty. The preregistered conjunction remains failed because the sparse-to-dense full-minus-early interval crosses zero. Internal summaries are also insufficient mediators: residual-norm change has $\rho = -0.02$ with prediction change; a late-layer orientation statistic reaches $\rho = 0.33$ and 0.70 across directions on a fresh universe, but its adjusted coefficients cross zero and it reduces the depth effect by only 9%.

A secondary forced-hop GOBLIN control gives zero failure when the operator omits adjacency ($k = 0$), 0.096 $[0.092, 0.100]$ at $k = 2$, and 0.084 $[0.077, 0.091]$ at $k = 4$. It therefore supports a second repeated-adjacency pathway, but the preregistered conjunction fails because sensitivity is not monotone in k.

C.3 RESIDUAL BOUNDARY AND COMPARISONS

The within-checkpoint reversal replicates across three universes (Table 8). The same fixed residual helps when GraphPFN's adapters are absent and becomes neutral or harmful at full depth.

Table 8: Residual accuracy gain by GraphPFN adapter depth. The last column is the paired depth-0 minus depth-12 gain.

Universe	Depth 0	Depth 12	Paired difference [95% CI]
4	$+0.031$ $[0.027, 0.036]$	-0.015 $[-0.019, -0.011]$	$+0.046$ $[0.039, 0.053]$
5	$+0.025$ $[0.019, 0.032]$	-0.005 $[-0.009, 0.000]$	$+0.030$ $[0.023, 0.037]$
9	$+0.059$ $[0.051, 0.067]$	-0.004 $[-0.008, 0.000]$	$+0.063$ $[0.054, 0.071]$

Table 9 places the reference update within the propagation family. The comparisons are descriptive and use the same cached predictions. All three preregistered superiority statements fail: plain propagation is competitive for SAMGPT, scale matching is not necessary for G2T-FM accuracy, and two steps outperform one where propagation helps. Scale matching mainly prevents severe calibration changes for G2T-FM and limits accuracy loss for MDGFM.

For the alignment check, the 25 cached conditions comprise 17 GraphPFN adapter conditions, two post-hoc architecture extensions, and six prospective cross-model conditions. One GraphPFN condition has an observed NLL interval containing zero; in every other condition the second-order prediction has the observed NLL sign (Table 10). Across all 25, predicted and observed NLL changes have Spearman correlation 0.985, whereas alignment and accuracy change correlate only 0.50. The paired alignment drop from depth 0 to depth 12 is 0.30 $[0.26, 0.33]$, 0.22 $[0.19, 0.25]$, and 0.26 $[0.24, 0.29]$ in universes 4, 5, and 9.

Table 9: Accuracy gain over the base prediction, pooled over two prospective universes and both extreme directions. C&S denotes its smoothing stage only.

Update	G2T-FM	SAMGPT	MDGFM
Reference residual	$+0.027$	$+0.014$	-0.004
Label propagation, one step	$+0.017$	$+0.018$	-0.024
C&S, two steps	$+0.012$	$+0.003$	-0.029
No scale matching, $\eta = 0.5$	$+0.027$	-0.002	-0.038
Residual, two steps	$+0.041$	$+0.017$	-0.012

Table 10: Coverage of the error-alignment sign check. Decisive conditions have observed NLL intervals excluding zero.

Condition family	Tested	Decisive	Correct sign
GraphPFN adapter conditions	17	16	16
GOBLIN and AnyGraph	2	2	2
G2T-FM, SAMGPT, and MDGFM	6	6	6
Total	25	24	24

The first prospective residual gate passes, but the stronger five-part headroom-gate replication remains failed because superiority over the ungated update was not established for MDGFM. Accordingly, neither the gate nor the label-dependent alignment diagnostic is presented as a deployment-time selector. Real-graph scope is reported separately in Appendix D.

D Real-Graph Evaluation

D.1 Paired thinning protocol

Real graphs cannot be densified without inventing edges, so we construct nested views by thinning. Each graph is partitioned without labels into node-disjoint context and query blocks. One uniform random variable is assigned to every observed edge, and increasingly dense views retain progressively larger prefixes of the same edge ordering. Node identities, features, and labels remain fixed within a block. Features are reduced by a label-blind truncated SVD and standardized; the width is selected without adapter outcomes when needed for an adjacency-blind headroom check. Inference uses the same pinned model interfaces as the synthetic benchmark, and intervals resample query blocks after averaging model seeds.

This construction tests sensitivity to removing observed edges, not densification beyond the recorded graph or transfer between independently collected datasets. PubMed is excluded before inference because of pretraining overlap. Flickr and Coauthor CS fail the minimum three-view requirement. SAMGPT rows on Amazon Computers and Photo are treated as descriptive because those graphs occur in its pretraining set.

D.2 Original labels

Table 11 reports the GraphPFN adapter-path contrast on eligible real graphs. High-homophily graphs with informative features show essentially no effect. Squirrel and arxiv-year exhibit a positive trace, but only at one to two accuracy points. Three Facebook100 graphs with real gender labels are again within one point. On Squirrel and arxiv-year, G2T-FM, SAMGPT, and GOBLIN have matched-density interactions between 0.00 and 0.007; only GraphPFN shows the small positive trace.

The flat results are not merely a ceiling artifact. Feature degradation and larger blocks keep the real-label contrast in the same few-point regime. Conversely, homophily and density ratio cannot explain the synthetic–real gap: within the calibrated generator, higher homophily increases the contrast and even a fourfold shift produces 0.15–0.29 of failure.

Table 11: Real-label GraphPFN depth-12 minus depth-0 transfer failure under paired thinning. Homophily is measured on the original graph.

Dataset	Edge homophily	Adapter-path contrast [95% CI]
Physics, Computers, Photo (pooled)	0.86–0.93	-0.001 $[-0.007, 0.005]$
Penn94	0.51	$+0.008$ $[0.002, 0.015]$
Cornell5	0.55	$+0.006$ $[0.002, 0.011]$
JohnsHopkins55	0.55	-0.001 $[-0.010, 0.008]$
Squirrel	0.22	$+0.020$ $[0.005, 0.034]$
arxiv-year	0.22	$+0.013$ $[0.005, 0.020]$

D.3 Holding real edges fixed and changing the label process

To test whether real topology or its relationship to the task suppresses the pathway, we keep real graph blocks and nested edges fixed but plant community labels and class-conditional features using the generator code. Table 12 contrasts these runs with the original labels. On three social graphs, planted labels restore 0.05–0.17 of adapter-path failure while the original labels remain within one point. Squirrel is the counterexample: even planted labels yield at most a two-point effect, including with larger blocks. Thus label–topology alignment explains the gap on some real edge sets, but the edge structure also determines whether GraphPFN's adapters can extract that signal.

Table 12: Adapter-path contrast on fixed real edge sets under original and planted labels. Ranges span the reported planted-homophily arms.

Edge set	Original labels	Generator labels
Penn94	$+0.008$	$+0.112$ to $+0.171$
Cornell5	$+0.006$	$+0.050$
JohnsHopkins55	-0.001	$+0.108$
Squirrel	$+0.020$	-0.001 to $+0.023$

These experiments bound rather than replace the synthetic benchmark. They show that the controlled pathway can occur on real edges when those edges carry the task labels, but do not establish its prevalence in naturally occurring graph prediction problems.