THE TRAINING-SET SHADOW: HOW MESSAGE PASSING BREAKS CONFORMAL COVERAGE ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learning models on graphs are often deployed with a reliability promise attached: conformal prediction gives each node a set of candidate labels that contains the true label a chosen fraction of the time, say 90%. We show this promise, while true on average, fails in a systematic and invisible way. Across seven standard benchmarks with few labeled nodes, the nodes sitting next to the training examples are the ones most often shortchanged, covered as little as 78% of the time under the most widely used scoring rule (adaptive prediction sets), even though the model classifies them more accurately than the rest of the graph, while faraway nodes receive more coverage than they need. A core of this failure persists under every scoring rule wherever the model makes confident mistakes, and the effect reappears, smaller, at much larger scale on a public citation network of over $160{,}000$ nodes. We call this blind spot the *training-set shadow*, and we show how to see it: report coverage as a function of each node's distance to the training set, a profile whose one-number summary we call the *shadow gap* and which costs one graph traversal plus a labeled holdout to compute. Controlled experiments pin the cause on message passing, the mechanism by which graph neural networks blend each node's features with its neighbors': a feature-only model on identical data shows no shadow, moving the training nodes moves the shadow with them, and the effect switches on the moment a single round of propagation is added. We prove where the leak lives: near the training set, the network's representations contain copies of the very examples it was trained on, exactly for linear networks and in the infinite-width limit for nonlinear ones, where the resulting kernel reproduces the shadow with no training at all; that this containment inflates confidence is established experimentally. The repair is simple and carries a finite-sample proof: calibrate separately within each band of distance to the training set, which restores the target coverage in every band at a small cost in prediction-set size, whereas calibrating on other graph structure such as communities, though exactly valid on its own terms, leaves the shadow untouched. Practitioners should report the coverage-by-distance profile alongside the marginal average, and calibrate by position, not only globally.

1 INTRODUCTION

In semi-supervised node classification, a small set of nodes carries labels and a graph neural network (GNN) spreads that supervision along edges to predict labels for the rest. Conformal prediction (Vovk et al., 2005) can wrap any trained GNN in this transductive setting (Huang et al., 2023; Zargarbashi et al., 2023; Clarkson, 2023), converting the model's confidence into a *prediction set* for each node: a set of candidate labels that contains the true label with probability at least $1 - \alpha$.

That guarantee is *marginal*, an average over nodes, so it can hold while specific groups are covered far less often; exact per-node conditional coverage is impossible without distributional assumptions (Barber et al., 2021). We do not claim the marginal guarantee fails (it holds throughout); our question is what the average conceals. When coverage is 90% on average, which nodes fall below it, and why? The field does not yet know how the marginal guarantee distributes coverage over a 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.

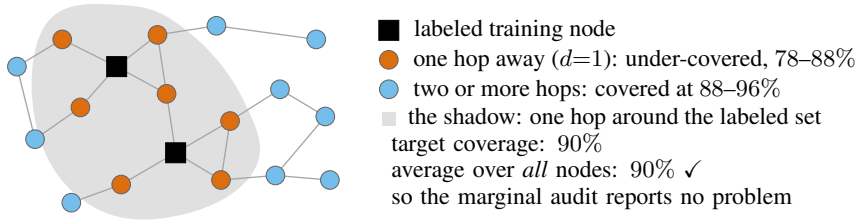

Figure 1: The training-set shadow, in one picture. Squares are the labeled training nodes; every other node is colored by how reliably conformal prediction covers it at a 90% target. Message passing hands each node its neighbors' features, so a node one hop from a training node (orange, shaded region) carries a copy of the very features the model was trained to fit. The model is overconfident there: those nodes get the smallest prediction sets, the sets miss most often, and coverage drops to 78–88%, while farther nodes (blue) are covered at or above target. The graph-wide average is still 90%, so a marginal audit sees nothing wrong. The shadow moves with the labeled nodes and disappears without message passing (§5).

We show the answer has a specific, reproducible shape: **the one-hop neighbors of the training set are systematically under-covered**, by up to 12 percentage points below target on all seven sparse-label benchmarks under the standard randomized adaptive prediction sets (APS) score, with a score-independent confident-error core on the graphs where the model errs confidently, while distant nodes are over-covered so that the average still comes out right. We call this the *training-set shadow*. Figure 1 illustrates the idea and Figure 2 shows the measurements. A marginal audit cannot see the shadow, and it falls on the nodes where graph-based semi-supervised learning claims its advantage. In representation-learning terms, the shadow is about what message passing writes into node embeddings: near the supervision the representation contains the training data itself (Results 2, 4), and calibrated uncertainty inherits that geometry.

This finding is a warning: a deployed GNN reporting "90% coverage" may cover the nodes next to its labeled data only 78–88% of the time. It motivates both a diagnostic worth reporting by default (the coverage-by-distance profile and its shadow gap, Algorithm 1) and a repair (per-distance-band thresholds with an exact guarantee, no retraining).

Our contributions:

- **The phenomenon (§4).** On $7/7$ sparse-label benchmarks, conformal prediction with the widely used APS score under-covers the training set's one-hop neighborhood (PubMed: 77.9% at a 90% target); the dip replicates on ogbn-arxiv (-2.7pp at 2% labels, $t{=}-15.0$) with a monotone density dose-response, and vanishes under dense labels (Table 2).
- **The cause is message passing (§5).** A multilayer perceptron (MLP) sharing features, splits, and conformal randomization shows no dip anywhere; the shadow moves with a re-sampled mask, survives within degree quartiles and tuned early-stopped training, switches on at one propagation step at fixed capacity, and tracks neighbor-mixing strength across architectures.
- **Not an accuracy artifact (§4).** On $5/7$ datasets the under-covered stratum is the *most* accurately classified (Coauthor-Physics: 96.0% accuracy, 85.2% coverage); the shadow is overconfidence, not difficulty.
- **A mechanism that explains and predicts (§6).** A coverage-transfer result predicts which scores show the shadow; a closed-form kernel (Result 2) locates the leak for linear GNNs; and coverage misses concentrate on label-disagreeing training edges, a penalty message passing roughly triples over the MLP.
- **A fix with an exact guarantee (§7).** Per-hop-distance Mondrian calibration restores every stratum to ≈90% exactly (Result 3) at 0.1–1.2 extra labels per set at $d{=}1$; diffusion adaptive prediction sets (DAPS) soften but do not repair the shadow, and other structure-aware methods guarantee only marginal or approximate coverage.

Two supporting studies complete the picture: the choice of nonconformity score decides the fairness verdict on GNNs, flipping published conclusions on their own benchmarks (§C.1), and per-class

coverage gaps are driven by classifier accuracy rather than homophily (§C.2), which is what Result 1 requires.

2 Related Work

Conformal prediction on graphs. Huang et al. (2023) prove transductive validity and learn a topology-aware correction, Zargarbashi et al. (2023) introduce the diffusion score DAPS, Clarkson (2023) handle the inductive case over random local patches, and Zargarbashi & Bojchevski (2024) give guarantees for evolving graphs; recent methods localize calibration with graph kernels (Baghershahi et al., 2026) or communities (Zhang et al., 2025). None examines coverage relative to the position of the training set, and no prior work identifies the train-proximity failure or uses the exactness of train-proximity strata (Result 3).

Conditional coverage. Exact conditional coverage is impossible without distributional assumptions (Barber et al., 2021); achievable forms are finite partitions (Vovk et al., 2005; Ding et al., 2023) and finite-dimensional shift classes (Gibbs et al., 2025; Tibshirani et al., 2019), with localized CP (Guan, 2023) and multivalid/multigroup guarantees (Bastani et al., 2022; Jung et al., 2023) the non-graph ancestors of the locality and group axes. We instantiate the achievable form with graph-position partitions.

Position, calibration, and evaluation in GNNs. Localized training sets induce distribution shift between labeled nodes and the rest (Zhu et al., 2021), accuracy is higher near labeled nodes (Han et al., 2023), generalization decays with distance (Ma et al., 2021), and evidential uncertainty grows with it (Stadler et al., 2021); the shadow is this literature's reverse side, since the accurate, low-uncertainty near-label nodes are the ones that *lose* coverage. GNNs are globally *under*-confident (Wang et al., 2021) (yet locally overconfident near supervision) and calibration error is lowest near the training nodes (Hsu et al., 2022), which explains why the shadow went unnoticed: the same confidence inflation that flatters calibration error removes the slack coverage needs (Result 1); the two lenses reached opposite verdicts, and for a deployed conformal system the coverage verdict binds. Conformal fairness audits (Romano et al., 2020a; Wu et al., 2025; Vadlamani et al., 2025; Ding et al., 2023) produce verdicts we show are score-dependent (§C.1); in the spirit of Shchur et al. (2018) and Platonov et al. (2023), we argue graph conformal evaluation should report coverage by position and score, not only marginally.

3 Setup

Conformal prediction. A classifier's probabilities are often wrong as frequencies: nodes labeled "95% confident" may be right far less often. Conformal prediction sidesteps trusting them by outputting a *set* of labels per node, sized so the true label lands inside a chosen fraction of the time, say 90%. The recipe is a ranking trick: on held-out labeled *calibration* nodes, score how implausible each true label looks to the model (a *nonconformity score*, e.g. one minus its assigned probability), take the 90th percentile as threshold $\hat{q}$, and give each test node every label scoring below $\hat{q}$. Easy nodes get small sets, hard nodes large ones, and the 90% rate holds by construction, requiring only that calibration and test nodes are interchangeable. The catch, and our subject, is that the 90% is an average: it says nothing about *which* nodes get covered, and on graphs the misses are systematic.

Transductive split conformal. We work with a fixed graph $G = (V, E)$ with features and labels (x_v, y_v), a fixed training set $T \subset V$ (plus a validation set), and the remaining pool P; the validation set is excluded from the pool and unused except in the tuned-training arm (all other arms share one fixed recipe so contrasts stay paired; App. F, N). A GNN is trained on T only and outputs a softmax vector for every node. The pool is split uniformly at random into a calibration half and a test half. Given a nonconformity score $s(x, y)$, the threshold $\hat{q}$ is the $\lceil (1-\alpha)(n+1) \rceil$-th smallest calibration score, and the prediction set for a test node is $\{y : s(x, y) \leq \hat{q}\}$. The trained model never sees calibration or test labels, and the calibration/test assignment is a uniformly random split, so the scores are exchangeable and marginal coverage of at least $1 - \alpha$ holds despite the edges (Huang et al., 2023).[1] We use $\alpha = 0.1$ and four standard scores: the least ambiguous set-valued classifier

[1]The inductive setting, where test nodes are unseen during training, breaks this exchangeability. App. P shows the shadow persists there and the fix still restores per-stratum coverage.

LAC ($s = 1 - \hat{p}_y$; Sadinle et al., 2019), randomized **APS** (Romano et al., 2020b), its regularized variant **RAPS** (Angelopoulos et al., 2021), and the graph-native diffusion score **DAPS** (Zargarbashi et al., 2023).

Position strata. For a pool node v, let $d(v, T)$ be the hop distance to the nearest training node. We group the pool into strata $d \in \{1, 2, 3, \geq 4\}$. The key property is that $d(\cdot, T)$ depends only on (G, T), both fixed *before* the calibration/test split is drawn; this is what makes the strata admissible for calibration (§7). Unless a table states otherwise, a "dip" is d=1 coverage minus the unweighted mean of the farther strata; the ogbn-arxiv table pools farther test nodes instead, since its stratum sizes differ by orders of magnitude. Datasets, models, and protocol details are in Appendices E and F. Models are two-layer graph convolutional networks (GCNs, width 64) unless stated otherwise, with a feature-only MLP of the same width and training recipe as the no-message-passing control.

4 The Training-Set Shadow

Table 1 and Figure 2 (left) report coverage by distance stratum on the seven standard sparse-label benchmarks (5 seeds). Under a global threshold with APS, the d=1 stratum is under-covered on all seven (77.9% on PubMed, 83.7% on CiteSeer, against a 90% target) while the strata at $d \geq 2$ mostly sit at or above target (88–96%): the marginal guarantee is paid for by over-covering distant nodes. Amazon-Computers is the one exception: its small, hard far strata degrade for GCN and MLP alike, a shared-difficulty effect separate from the shadow, so there the d=1 dip shows only against the target and against the dominant d=2 stratum (85.8 vs 93.6%), with the size-weighted marginal at 91.3%. The dip is not noise: the strata are large (the d=1 column alone contains 141 to 1,541 test nodes), the paired-by-seed contrast between d=1 and the farther strata is significant on 6/7 datasets ($|t| = 4.1$–16.7), and Amazon-Computers is significant against its dominant d=2 stratum ($t = -17.6$).

The audit we recommend is the full coverage-by-distance profile, summarized by its worst-stratum gap (target coverage minus the minimum stratum coverage). On all seven benchmarks the minimum stratum is d=1, so the summary reduces to the *shadow gap*, target minus d=1 coverage: 12.1 percentage points on PubMed, 6.3 on CiteSeer, 4.1–5.8 on the Amazon and Coauthor graphs, and 1.6 on Cora (Table 1). These are single-mask numbers; over ten fresh training masks per dataset (Table 7) the dip is negative on 10/10 masks on five datasets (9/10 on Amazon-Photo) with the paired MLP at zero, and mask-averaged magnitudes shift both ways: PubMed to -5.4 ± 2.7, CiteSeer to -15.4 ± 2.9 (deeper than its public-split value; mask-averaged gap 12.8pp). At scale the profile matters: on ogbn-arxiv the worst stratum is the far one (below), so the profile flags both failure modes while a d=1-only number would not. Algorithm 1 computes the profile.

The shadow at scale. The effect is not a small-graph artifact. On ogbn-arxiv (169,343 nodes, 2.3M undirected edges, 40 classes), re-sampling stratified training sets at 2% and 5% density reproduces the pattern (Table 2): d=1 covers 87.9% and 88.4% against 91.6–91.8% at d=2, a dip of -2.7 ± 0.4 and -2.3 ± 0.2pp against the pooled farther strata ($t = -15.0$ and -23.2; the d=1 stratum alone holds 15,812–28,778 test nodes), and the accuracy flip replicates: d=1 is the *most* accurately classified stratum (73%, falling to 51% at $d \geq 4$) yet sits below target. The far strata also degrade (85% at $d \geq 4$), tracking their low accuracy, the same shared-difficulty effect as Amazon-Computers; at this scale the far stratum, not d=1, is the profile's worst entry, which is why we recommend the full profile over any single-stratum number. The shadow remains the anomalous entry: d=1 is under-covered while being the *most* accurate stratum, whereas the far deficit tracks accuracy. Distance-Mondrian restores every stratum, near and far, to ≈0.90 in every arm. The density dose-response holds at scale: the dip attenuates across 2/5/10/25% labels ($-2.7 \to -2.3 \to -2.3 \to -1.9$pp), and under the official 54% split, where d=1 is 93% of the pool, no significant dip remains (-1.1 ± 1.0pp). The dip magnitude is smaller than on the citation graphs, consistent with arxiv's higher-degree neighborhoods diluting any single training node's influence, but its sign and repair are unambiguous at this sample size.

Low accuracy is not the cause. A natural objection is that nodes near the labels might simply be harder to classify. The opposite holds. On 5/7 datasets the d=1 stratum is *more* accurately classified than the farther strata on average (accuracy-gap column of Table 1), and on every dataset

Table 1: The training-set shadow (GCN, APS, 5 seeds, target 90%). The d=1 stratum is under-covered on all seven benchmarks, yet its *accuracy* is usually the highest (acc. gap: accuracy at d=1 minus the farther-strata mean; positive means the under-covered nodes are the more accurate ones). The MLP control shows no shadow; distance-Mondrian (§7) repairs it. *: dip significant (paired-by-seed t, $p < 0.05$). †: Amazon-Computers' farther-strata mean is distorted by its collapsing far strata; vs its dominant d=2 stratum the dip is -7.7 ± 1.0pp ($t = -17.6$).

		coverage (%)		MLP	acc. gap	LAC	dist-Mond.
Dataset	$n_{d=1}$	d=1	$d \geq 2$ (mean)	d=1	at d=1 (pp)	d=1	d=1
Cora	203	$88.4 \pm 2.3^*$	93.1	89.9	+4.0	90.6	91.1
CiteSeer	141	$83.7 \pm 1.9^*$	92.7	93.0	+3.7	85.2	90.5
PubMed	142	$77.9 \pm 3.8^*$	89.7	88.5	−5.6	78.8	91.1
Amazon-Photo	1,261	$85.9 \pm 1.0^*$	94.3	91.7	+2.9	90.6	90.1
Amazon-Computers	1,541	$85.8 \pm 1.5^\dagger$	85.8	90.6	+10.7	87.6	90.5
Coauthor-CS	1,084	$84.2 \pm 1.5^*$	90.2	90.7	+0.2	92.9	90.4
Coauthor-Physics	556	$85.2 \pm 2.3^*$	90.2	90.0	+2.0	94.7	89.7

Table 2: The shadow on ogbn-arxiv (169,343 nodes; GCN, APS, 5 seeds, target 90%). Sparse training sets show the d=1 dip and distance-Mondrian repairs all strata; the official dense split shows no significant dip, the scope condition at scale. Dip is d=1 vs pooled farther strata, paired by seed. †: under the official split the far strata are near-empty remnants (mean $n = 166$ and 21): their global-threshold coverage estimates are noisy and accuracy-driven, and the per-stratum Mondrian thresholds are conservative at such counts (upper end of the Mondrian range).

		coverage (%) by hop distance					
Training set	n_{train}	d=1	d=2	d=3	$d \geq 4$	dip (pp)	Mond. range
2% stratified	3,387	87.9	91.6	90.1	85.4	$-2.7 \pm 0.4^*$	89.8–90.3
5% stratified	8,466	88.4	91.8	88.8	84.2	$-2.3 \pm 0.2^*$	89.9–90.1
10% stratified	16,936	89.0	92.3	87.9	84.3	$-2.3 \pm 0.1^*$	89.8–90.9
25% stratified	42,337	89.6	92.3	85.5	83.4	$-1.9 \pm 0.1^*$	89.9–90.4
official (54%)	90,941	89.9	91.7	$86.5^\dagger$	$82.0^\dagger$	-1.1 ± 1.0	$89.9–94.2^\dagger$

the coverage dip is larger than the accuracy dip. Coauthor-Physics is the cleanest case, pairing 96.0% accuracy at d=1, tied for the highest of any stratum, with 85.2% coverage, the lowest. On PubMed and Amazon-Computers accuracy does dip at d=1, so part of their shadow is accuracy-driven, but even there the coverage dip is larger. The prediction sets themselves show the overconfidence: the d=1 stratum receives the *smallest* sets on every dataset (Cora 1.61 labels vs 2.11–2.72 farther out; Coauthor-Physics 0.96 vs 1.05–1.40) while being covered least, and small sets that miss their targets are the signature of overconfidence. Near its supervision, the GNN is right more often but sure of itself far too often.

The score decides where the shadow appears, as predicted. Under the LAC score the shadow persists where the d=1 stratum makes *confident errors* (PubMed 78.8%, CiteSeer 85.2%) and vanishes or inverts where that stratum is accurate but overconfident (Cora 90.6%, Coauthor-CS 92.9%), the pattern consistent with Result 1: LAC's miscoverage tends to the stratum's error rate, while randomized APS additionally penalizes overconfident *correct* predictions. Decomposing the APS shadow directly (App. O, Table 16): a confident-error core survives with empty sets disallowed (-6.2/-8.8pp on CiteSeer/PubMed, matching LAC), the Coauthor shadows are carried by the randomized empty-set channel (empty sets concentrate at d=1: 12.5–12.7%), and de-randomizing is no escape from the geometry: deterministic APS widens the d=1-vs-farther disparity to -11 to -23pp, chiefly by pushing far strata toward ceiling coverage while d=1 errors remain (a redistribution as well as a validity statement). Both channels are positional; the profile audit catches both.

Scope: sparse supervision plus confidence concentration. On the five 50%-labeled heterophily benchmarks of Platonov et al. (2023) no stratum falls below target (Amazon-ratings' one significant gap is far-stratum *over*-coverage against an at-target d=1). Density alone does not explain

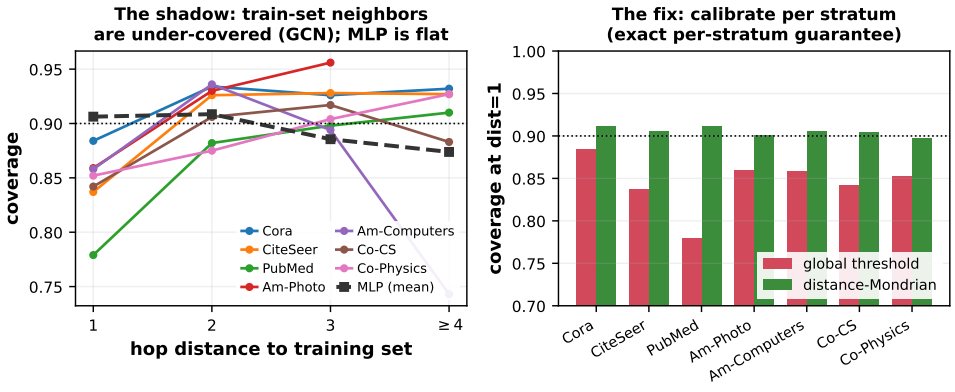

Figure 2: The training-set shadow, measured. *Left:* coverage vs. hop distance to the training set (GCN, APS). All seven sparse-label benchmarks dip at $d{=}1$; the MLP control (dashed) is flat. *Right:* distance-Mondrian calibration restores the $d{=}1$ stratum to target on every dataset, with an exact per-stratum guarantee (Result 3).

this: sparse 2–10% masks on Roman-empire, Amazon-ratings, and Minesweeper still produce no shadow (dips -0.8 to $+0.2$pp, all n.s.), because propagation never concentrates confidence near the labels there ($d{=}1$ accuracy equals far accuracy, 15.3 vs 15.1% on Roman-empire; the tuned grid of App. N lifts accuracy only to 17.6% and the null persists), so Result 1 has nothing to transfer. Whether heterophily-specialized architectures that actually learn these graphs cast a shadow remains open. On homophilous graphs density is the dial: deepest at 2% (-12.4pp on Cora), shrinking as labels densify, with a residual possible (-4.1pp on Cora at 50%). The scope is mechanistic: sparse supervision *and* confidence concentrated around it.

5 The Cause: Message Passing

Four controls isolate the cause.

Control 1: remove message passing. The MLP column of Table 1 uses the same features, training nodes, calibration/test splits, and conformal randomization u as the GCN; only the architecture differs. Its $d{=}1$ coverage is 88.5–93.0%, with no dip on any of the seven datasets (both architectures degrade on Amazon-Computers' small, hard far strata, the shared-difficulty effect noted above). Whatever creates the shadow travels along edges.

Control 2: move the training set. If the labeled set casts the shadow, moving the labels must move the shadow. Re-sampling the training mask (same per-class counts, new nodes), retraining, and re-stratifying by distance to the new mask, the shadow follows: the new $d{=}1$ stratum covers at 78.4% on CiteSeer (vs 92–94% farther), 85.9% on Cora, 83.8% on PubMed, 85.4% on Coauthor-CS, and dips by -5.8 ± 0.6pp on Coauthor-Physics ($t = -20.7$), while the MLP shows no dip in any of these cases (the Amazon graphs' far-strata collapse dominates their resampled arms; Table 7 extends this control to ten masks per dataset). The shadow is a property of where the supervision sits, not of any fixed set of nodes.

Control 3: degree. Training-set neighbors have much higher degree than other nodes (median 35.0 vs 15.4 on Amazon-Photo), and GNN confidence varies with degree, so the shadow could be degree composition in disguise. Comparing $d{=}1$ against farther nodes *within* rank-based degree quartiles rules this out: the gap (pooled over farther nodes within bins, as this control requires) stays negative and significant on 7/7 datasets ($|t| = 3.9$–19.7). On the citation and co-authorship graphs the within-bin gap essentially equals the raw gap (Cora -4.4 vs -4.7pp); on the Amazon graphs degree explains roughly half, with the rest still highly significant. Comparisons use only degree bins where both groups have ≥ 20 nodes.

Control 4: turn propagation on and off at fixed capacity. A decoupled architecture (APPNP, approximate personalized propagation of neural predictions; Gasteiger et al., 2019) separates propagation from capacity: a fixed two-layer MLP produces logits, then parameter-free personalized-PageRank runs K steps, so K is a pure propagation dial and $K{=}0$ reproduces the MLP control bit for bit. The shadow is off at $K{=}0$, on at $K{=}1$ (significant on $6/7$ datasets, $5/7$ after within-family BH; PubMed -9.0pp), and saturates for $K \geq 2$ (Table 9), while accuracy moves the *opposite* way ($d{=}1$ accuracy on Cora rises $0.63 \rightarrow 0.85$). The input-side dial agrees: pre-propagated features $S^K X$, a scalable inception graph network (SIGN) featurization (Result 2's with a nonlinear head), fed to the same fixed MLP switch the shadow on at $K{=}1$ ($6/7$; Table 10). The mechanism is pinned from both directions at fixed capacity: propagation alone produces the shadow. (Amazon-Computers inverts at all K, its weak-mixing family pattern.)

Architecture scope. The dip is significant on $7/7$ benchmarks for GCN, $4/7$ for the graph attention network (GAT), $2/7$ for GraphSAGE (Table 20): SAGE's root weight and GAT's attention dilute neighbor mixing and thereby the leaked training signal, the ordering the mechanism predicts. It extends beyond message passing in the strict sense: the spectral Chebyshev network (ChebNet) dips on $6/7$ ($5/7$ significant), while a preliminary GraphGPS study suggests all-to-all attention *moves* conditional miscoverage off $d{=}1$ rather than eliminating it (App. Q). Which architectures are affected on which graphs is not predictable in advance; audit any propagating architecture.

Training-recipe scope. The shadow is not a fixed-recipe artifact: in a Shchur-style tuned arm (per-(dataset, seed) grid, validation early stopping, App. N) the dip survives on $4/7$ benchmarks at equal or larger magnitude (Cora -6.5, CiteSeer -8.0, PubMed -11.8, Coauthor-CS -7.4pp), attenuating to n.s. on Amazon-Photo and Coauthor-Physics. The regime does matter: on ogbn-arxiv at 2% labels a standard tuned recipe (3-layer, BatchNorm, early-stopped) shows *no* $d{=}1$ shadow (91.3–91.9%), leaving only the accuracy-tracking far deficit. Training choices decide where conditional coverage fails; the profile audit is the recipe-independent prescription.

6 The Mechanism: From Overconfidence to Miscoverage

The mechanism needs two ingredients, a reason why coverage fails on confident errors and a reason why GNN confidence concentrates around the training set.

Result 1 (Coverage transfer)**.** *Fix any score s and global threshold $\hat{q}$, and assume $\hat{q}$ stays interior, $\hat{q} \in (\epsilon, 1-\epsilon)$ for some $\epsilon > 0$ (the stratum may peak while the overall calibration distribution keeps $\hat{q}$ interior). For any stratum S of the test nodes, the miscoverage of S equals* $\Pr(s_{\text{true}} > \hat{q} \mid S)$*. In the peaked-prediction limit on S ($\hat{p}_{\text{top}} \rightarrow 1$): for LAC, the stratum miscoverage tends to the stratum's error rate; for randomized APS it tends to* $\text{err}(S) + (1 - \text{err}(S))\ \Pr(u > \hat{q})$*, i.e. an additional uniform penalty on overconfident* correct *predictions.*

Why this holds. The first claim is the definition: a test node is covered exactly when its true-label score is below the global quantile. For LAC, $s = 1 - \hat{p}_y \rightarrow 0$ on peaked-correct nodes and $\rightarrow 1$ on peaked-wrong nodes, so the stratum's mass above $\hat{q} \in (\epsilon, 1-\epsilon)$ is its error rate. For randomized APS, a peaked-correct node has $s \approx u$, covered only when $u \leq \hat{q}$; a peaked-wrong node has $s \approx 1$. □

The APS clause formalizes the known small-set behavior of the randomized score (Romano et al., 2020b; Angelopoulos & Bates, 2023); the content is its transfer, with confident errors, into *positional* coverage on graphs.

Result 1 is stated in the peaked-prediction limit and serves as the interpretive account rather than a premise of any empirical claim; empirically the transition into the regime is smooth (coverage is nominal for $\hat{p}_{\text{top}} \in [0.4, 0.6]$ and decays to 0.78 at $\hat{p}_{\text{top}} \geq 0.95$, one 0.8pp bin-to-bin uptick aside, even as accuracy rises to 0.99, with $d{=}1$ nodes over-represented in the high-confidence bins; App. M, Table 14). A non-asymptotic treatment is left to future work.

In words: a well-calibrated model lets conformal prediction absorb difficulty into set size; an overconfident one turns errors, and under APS even overconfident correct answers, into coverage failures. The graph-specific ingredient is that one hop from T a node's representation aggregates the very training features the loss fit, biasing its logits toward the training neighbors' labels: agree-

ment (the homophilous majority) yields peaked-correct predictions that randomized APS misses with probability $\approx 1 - \hat{q}$, disagreement yields the confident errors every score misses. For linear GNNs the bias has a closed form:

Result 2 (Label bleed in linear GNNs). *Let $Z = S^L X$ be an L-layer linear GNN (a simple graph convolution, SGC) with propagation matrix S (normalized adjacency with self-loops), trained by ridge regression on the training rows, $W^\star = \arg\min_W \|Z_T W - Y_T\|_F^2 + \lambda \|W\|_F^2$ with $\lambda > 0$. Then every prediction is a training-label smoother, $f(v) = \sum_{u \in T} k(v,u)\, y_u$ with $k(v,u) = z_v \,(Z_T^\top Z_T + \lambda I)^{-1} z_u^\top$, and $z_v = \sum_w S^L(v,w)\, x_w$ contains the training node u's own features x_u with weight $S^L(v,u)$, which is nonzero exactly when u is within L hops of v and vanishes for the MLP ($S = I$, $v \neq u$). (Proof: App. H.)*

Read plainly: within L hops of a training node u, a node's representation literally contains a copy of u's features, the ones the loss was fit to, a term no feature-only model has. When $y_v = y_u$ this sharpens a correct prediction; when $y_v \neq y_u$ it pushes a confident wrong one, which by Result 1 becomes miscoverage. The smoother form places message passing in the lineage of label propagation (Zhu et al., 2003). To be precise: containment is what Results 2 and 4 prove; that the contained features inflate confidence is established by the controls and the microtest, not by a theorem.

Beyond linearity. In the infinite-width neural tangent kernel (NTK) regime, under the standard zero-initial-function device, the trained nonlinear ReLU GCN is exactly a training-label smoother whose kernel carries the same L-hop self-feature containment, vanishing for $S{=}I$; APPNP's kernel follows by an exact identity (Result 4, App. L). This is testable with no training: kernel ridge with the exact GCN-NTK reproduces the shadow on the citation graphs (CiteSeer -7.2pp, significant; Cora -2.8pp, raw $p < 0.05$ only), the $S{=}I$ kernel shows none, the APPNP-NTK family shows the dose-response in K, and the finite-network dip is stable in width ($16 \to 1024$). The Amazon graphs mark the boundary: their exact kernels show no dip although trained networks dip at every width, so there feature learning, which the lazy regime does not capture, drives the shadow. The finite-width nonlinear case rests on the controls of §5 and the microtest below.

Microtest: the misses sit on label-disagreeing training edges. The leakage account predicts that within $d{=}1$, under-coverage should concentrate on nodes none of whose adjacent training nodes shares their label. It does, on $6/7$ datasets: agreeing $d{=}1$ nodes are covered at 86–95% while label-disagreeing ones fall to 45–79% (PubMed 45.3%, CiteSeer 51.4%) at accuracies of 9–68%, and the MLP arm shows message passing roughly triples this penalty (-44 to -47pp for the GCN vs -14 to -17pp for the MLP on CiteSeer and PubMed). The one exception, Coauthor-CS, isolates the second component: its agreeing group pairs 97% accuracy with 84% coverage, the randomized-APS penalty on peaked-correct predictions, matching its missing LAC shadow. Full details in App. B.

7 The Fix: Position-Conditional Calibration, Exactly Valid

Exact per-node conditional coverage is impossible without distributional assumptions (Barber et al., 2021); coverage over finite partitions is the achievable form (Vovk et al., 2005; Gibbs et al., 2025). On graphs the natural partition is by position relative to the supervision, and it comes with a guarantee:

Result 3 (Position-conditional validity). *Under the transductive exchangeability condition of Huang et al. (2023) (fixed graph, fixed training/validation split, model trained without calibration or test labels, uniformly random calibration/test split of the pool), let Π be any partition of the pool measurable with respect to (G, T), e.g. hop-distance strata $d(\cdot, T)$, degree bins, or communities. Mondrian split conformal with per-cell thresholds attains, conditionally on the realized per-cell calibration counts, for every cell c with n_c calibration nodes, $1{-}\alpha \;\le\; \Pr\big(y_v \in C(v) \,\big|\, v \in c\big) \;\le\; 1{-}\alpha{+}\frac{1}{n_c+1}$.*

In plain terms: group the nodes using only the graph and the training-label positions, run the calibration recipe once per group, and each group gets its own 90% guarantee, exactly.

Why this holds. Π is fixed before the calibration/test randomization, so, conditional on a cell, the split restricted to that cell is still uniformly random; scores within the cell are exchangeable, and the standard split-conformal quantile argument applies per cell (Vovk et al., 2005). □

(The upper bound assumes a.s.-distinct scores; the probability is over the split randomness for the fixed graph, training set, and trained model, as in Huang et al., 2023.)

The fix works, and what it costs. Distance-Mondrian restores every $d{=}1$ stratum to 89.7–91.1% on all seven datasets (Table 1; Fig. 2, right) and every stratum on ogbn-arxiv, leaving other strata at target. The constraint is per-cell calibration mass (below 10 nodes we fall back to the global threshold; cells can be pooled per Ding et al., 2023); instrumenting the fallback shows it never fires at realistic budgets (0 of 35 dataset×ratio cells across splits 20/80–80/20) and degrades to the global threshold's coverage when starved (App. J). The price lands where coverage was missing: $d{=}1$ sets grow by 0.1–1.2 labels while over-covered middle strata shrink, and averaged over scores the trade is free (worst-stratum violation $+0.05$–$0.07 \to \leq +0.008$ at unchanged mean set size; App. M).

Why existing methods do not fix it. Diffusion and localized-kernel approaches (Zargarbashi et al., 2023; Clarkson, 2023; Baghershahi et al., 2026) improve efficiency but guarantee only marginal validity. Community-Mondrian variants (Zhang et al., 2025) give exact cell-conditional coverage on *communities*, and ablating that conditioning variable with our own machinery shows the axis is what matters: Mondrian over Louvain communities is valid per community yet leaves the $d{=}1$ dip intact on 7/7 datasets at near-global magnitude, unchanged on a coarser merged partition with per-community thresholds active for every evaluated node (App. Q, Table 19). The neighbourhood-adaptive prediction sets (NAPS) method (Clarkson, 2023) leaves $d{=}1$ the minimum-coverage stratum on 7/7 while blanket-over-covering far strata at 1.1–1.7× distance-Mondrian's set size (Table 18); DAPS softens the shadow but does not repair it (PubMed $d{=}1$: $77.9 \to 81.2\%$, no per-stratum guarantee); class-conditional calibration fixes only its own axis. Position-Mondrian is exact on the distance axis with no new machinery.

Algorithm 1 (App. A) gives the complete recipe; both steps run on top of any deployed conformal pipeline, with no retraining.

8 The Wider Picture: Scores and Accuracy

The shadow is the sharpest instance of a broader pattern: on graphs, conformal coverage is redistributed by choices the literature treats as details (full studies in App. C). First, the score decides the fairness verdict (App. C.1): the GCN-vs-MLP per-class disparity verdict reverses sign across LAC/APS/RAPS on 4/11 benchmarks, and changing only the score flips two published GNN-conformal-fairness conclusions (CF-GNN's PubMed disparity: $22.8 \to 1.2$pp; Table 3). Second, accuracy, not homophily, sets per-class severity (App. C.2): disparity tracks accuracy in a degree-preserving rewiring design, and homophily has no effect at matched accuracy. Both are the empirical face of Result 1.

9 Limitations, Future Directions, and Conclusion

Limitations. Established for standard transductive split conformal on sparse-label benchmarks; strongest for GCN, weaker for GAT/GraphSAGE, strongly attenuated under dense labels (a residual ∼4pp can persist at 50%), and it survives tuned training on 4/7 benchmarks, though the arxiv-scale tuned recipe relocates the deficit off $d{=}1$ (App. N). CF-GNN's correction network and GraphLCP are not re-run as repair baselines. Result 1's peaked limit is an approximation; Results 2/4 prove containment only, leaving the feature-learning regime (dominant on Amazon) open.

Future directions. The shadow persists inductively (4/4 datasets, App. P) and distance-Mondrian still repairs it; the growing-graph setting, label propagation, tuned transformers, and label-injection methods such as Correct & Smooth (Huang et al., 2021) are natural next tests, as is whether confidence-thresholded self-training amplifies confident errors. Labeling and calibration budgets could be designed jointly. A quantitative containment-to-miscoverage theorem would complete the theory.

Conclusion. Marginal conformal validity on graphs hides a systematic redistribution: message passing makes GNNs overconfident around their supervision, so the training set's neighbors, often the most accurately classified nodes, are the least reliably covered. Calibrating within distance strata repairs this exactly; report coverage by position, not only marginally.

AI USE STATEMENT

The research questions, experimental designs, formal claims, and final conclusions of this work are the authors' own, and the authors take full responsibility for its entire content. Generative AI tools were used, under author direction, to assist with drafting and editing the manuscript text, implementing experiment and figure code, drafting the proofs of Results 1, 2, and 3, and literature search; they were not used to generate synthetic data (dataset cleaning, translation, and qualitative analysis are not applicable here). All such output was checked by the authors: the proofs, which compose standard arguments (split-conformal exchangeability, Mondrian conditioning, and the ridge-regression smoother identity), were verified by hand; experiment code was reviewed for correctness before every run and every reported number was traced to its raw result file; the central empirical finding was independently reproduced by a separately written implementation; and all bibliography entries were verified against their sources.

REPRODUCIBILITY STATEMENT

All experiments run on CPU from public datasets (Planetoid, Amazon, Coauthor, the heterophily suite of Platonov et al. (2023), and ogbn-arxiv) with fixed seeds; the protocol, hyperparameters, and stratum definitions are specified in §3 and App. F, every reported number traces to a result file produced by a named script, and code to reproduce all tables and figures is included in the supplementary material.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, 2019.

Peyman Baghershahi, Fangxin Wang, Debmalya Mandal, and Sourav Medya. GraphLCP: Structure-aware localized conformal prediction on graphs. *arXiv preprint arXiv:2605.08074*, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, 2022.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, 2019.

Jase Clarkson. Distribution free prediction sets for node classification. In *International Conference on Machine Learning (ICML)*, 2023.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Simon S. Du, Kangcheng Hou, Ruslan Salakhutdinov, Barnabás Póczos, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In *Advances in Neural Information Processing Systems*, 2019.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Haoyu Han, Xiaorui Liu, Feng Shi, MohamadAli Torkamani, Charu C. Aggarwal, and Jiliang Tang. Towards label position bias in graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kexin Huang, Ying Jin, Emmanuel Candès, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. Subgroup generalization and fairness of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mikhail Mironov and Liudmila Prokhorenkova. Revisiting graph homophily measures. In *Learning on Graphs Conference (LoG)*. PMLR, 2024.

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

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2), 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Mahalakshmi Sabanayagam, Pascal Esser, and Debarghya Ghoshdastidar. New insights into graph convolutional networks using neural tangent kernels. *arXiv preprint arXiv:2110.04060*, 2023.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Relational Representation Learning Workshop (R2L), NeurIPS*, 2018.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Aditya T. Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. A generic framework for conformal fairness. In *International Conference on Learning Representations (ICLR)*, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Longfeng Wu, Yao Zhou, Jian Kang, and Dawei Zhou. Bridging fairness and uncertainty: Theoretical insights and practical strategies for equalized coverage in gnns. In *Proceedings of the ACM Web Conference (WWW)*, pp. 4625–4634, 2025. doi: 10.1145/3696410.3714909.

Soroush H. Zargarbashi and Aleksandar Bojchevski. Conformal inductive graph neural networks. In *International Conference on Learning Representations*, 2024.

Soroush H. Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning (ICML)*, 2023.

Zheng Zhang, Jie Bao, Zhixin Zhou, Nicolo Colombo, Lixin Cheng, and Rui Luo. Residual reweighted conformal prediction for graph neural networks. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence (UAI)*, 2025.

Qi Zhu, Natalia Ponomareva, Jiawei Han, and Bryan Perozzi. Shift-robust GNNs: Overcoming the limitations of localized graph training data. In *Advances in Neural Information Processing Systems*, 2021.

Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, 2003.

A PRACTITIONER RECIPE

Assessing the flag. The coverage indicators on H share the single estimated threshold $\hat{q}$, so they are not independent Bernoullis and a binomial test is anticonservative. A valid test targets the null *position does not matter*: holding $\hat{q}$ and the audit coverage indicators fixed, permute the stratum labels $b(\cdot)$ among the audit nodes M times (e.g. $M{=}1{,}000$), recompute each stratum's coverage under every permutation, and take the per-stratum permutation p-value of the observed deficit, with Benjamini–Hochberg across the D strata. (Re-drawing the calibration/audit split instead would resample under the alternative and is not a test.) The audit's headline is the full per-stratum profile; the worst-stratum gap summarizes it, and on our seven benchmarks that stratum is always $d{=}1$.

B EDGE-LEVEL MICROTEST: FULL VERSION

The leakage account makes an edge-level prediction: within $d{=}1$, under-coverage should concentrate on nodes none of whose adjacent training nodes shares their label. It does, on $6/7$ datasets (5 seeds). Nodes at $d{=}1$ with an agreeing training neighbor are covered at 86–95%, while label-disagreeing ones fall to 45.3% on PubMed, 51.4% on CiteSeer, 69.1% on Amazon-Computers, and 72.9–78.6% elsewhere, with accuracies of 9–68%, the confident-wrong push made visible. The MLP arm shows how much of this is the graph's doing: on the same stratification the MLP's agree-vs-disagree gap is -13.8 to -17.3pp on the citation graphs, where the GCN's is -43.9 to

Algorithm 1 Shadow audit and distance-Mondrian calibration (drop-in, no retraining)

Require: graph G, training set T, trained model, labeled calibration set C, audit holdout H (labeled, disjoint from C), test nodes, score s, target $1-\alpha$, strata cap D (e.g. 4), minimum cell size $n_{\min}$ (e.g. $\max(10, \lceil (1-\alpha)/\alpha \rceil)$)

$d(v) \leftarrow$ hop distance from v to nearest node of T (one breadth-first search); $b(v) \leftarrow \min(d(v), D)$

Audit (report alongside the marginal coverage number):

$\hat{q} \leftarrow \lceil (1-\alpha)(|C|+1) \rceil$-th smallest of $\{s(x_v, y_v) : v \in C\}$

for stratum $b = 1, \ldots, D$ **do**

  report $\widehat{\mathrm{cov}}_b \leftarrow \mathrm{mean}\{\, \mathbf{1}[s(x_v, y_v) \le \hat{q}] : v \in H,\ b(v) = b \,\}$

end for

flag any stratum b whose $\widehat{\mathrm{cov}}_b$ falls significantly below $1-\alpha$ (permutation test below); the d=1 stratum is the shadow

Fix (position-Mondrian; exact per-stratum coverage by Result 3):

for stratum b with $n_b = |\{v \in C : b(v) = b\}| \ge n_{\min}$ **do**

  $\hat{q}_b \leftarrow \lceil (1-\alpha)(n_b+1) \rceil$-th smallest calibration score in stratum b

end for

for sparse strata, pool adjacent strata or fall back to $\hat{q}$ (per-cell guarantee then reverts to marginal); note $n_{\min}$ near $\lceil (1-\alpha)/\alpha \rceil$ makes the per-cell threshold the cell maximum and hence conservative, so $n_c \gtrsim 25$–50 is advisable at α=0.1

prediction set for test node v: $\{y : s(x_v, y) \le \hat{q}_{b(v)}\}$

-47.3pp, so these nodes are moderately hard even without the graph, but message passing roughly triples their penalty (Cora is the mild case, -19.1 vs -16.2). The exception is instructive: on Coauthor-CS the *agreeing* group pairs 97% accuracy with 84% coverage, so that shadow is carried by the randomized-APS penalty on peaked-correct predictions, Result 1's second term, matching Coauthor-CS's missing LAC shadow. The microtest thus separates the shadow's two components, confident errors versus the overconfidence penalty, dataset by dataset.

C Supporting studies: scores and accuracy (full versions)

This appendix reports in full the two studies summarized in §8. The shadow is the sharpest instance of a broader pattern: on graphs, conformal coverage is redistributed by choices the literature treats as details. Two supporting studies map that wider pattern.

C.1 The nonconformity score decides the fairness verdict

Holding the model, the data, and the target coverage fixed, standard scores disagree about which model is fairer. On 11 of the 12 benchmarks (Coauthor-Physics excluded from this study) the per-class coverage-disparity verdict between a GCN and an MLP reverses sign across LAC/APS/RAPS on $4/11$ datasets under deployable split conformal, and on $6/11$ under a matched-coverage diagnostic that equalizes overall coverage first (the diagnostic uses test labels and is not a deployable method). The effect survives GAT and GraphSAGE backbones, DAPS, three disparity metrics, and Benjamini–Hochberg correction (App. R); off graphs the score-dependence appears in magnitude but not in sign (App. G). Reproducing two published GNN-conformal-fairness results on their own datasets and metrics, and changing only the score, flips their conclusions. The Conformalized GNN work (CF-GNN) of Huang et al. (2023) reports "satisfactory conditional coverage over network features" under APS, but the degree-conditional disparity that looks large under APS is small under an equally standard alternative (Table 3; PubMed $22.8 \to 1.2$pp), and the model ranked fairest by the group ΔCoverage metric of Wu et al. (2025) changes with the score on $4/5$ of their datasets ($2/5$ under the APS$\to$LAC swap alone). We reproduce protocols and metrics, not CF-GNN's correction network; the claim is that the reported quantities are score-fragile. On three datasets with genuine demographic attributes (NIFTY Credit, German, Bail) no score yields a significant group-coverage reversal ($0/3$); in our experiments the instability is structured by class and position, not demographics.

Table 3: Changing only the nonconformity score changes published conclusions on their own datasets: CF-GNN's (Huang et al., 2023) degree-conditional coverage disparity (GCN, pp; lowest per row in **bold**).

	APS	LAC	RAPS	DAPS
Cora	14.9	8.4	11.9	**4.6**
CiteSeer	9.8	6.8	6.9	**6.5**
PubMed	22.8	**1.2**	9.4	4.8
Coauthor-CS	12.6	6.7	12.3	**2.2**
Coauthor-Phys.	35.7	**15.0**	33.4	31.2

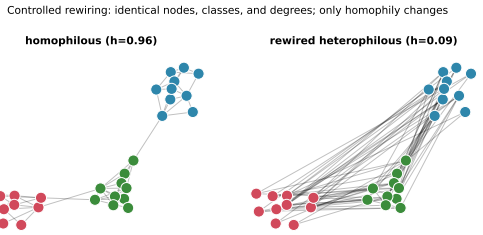

Figure 3: The controlled rewiring behind §C.2: two graphs with *identical* nodes, class labels, and per-node degrees, differing only in edge homophily (double-edge swaps convert same-class edges to cross-class edges).

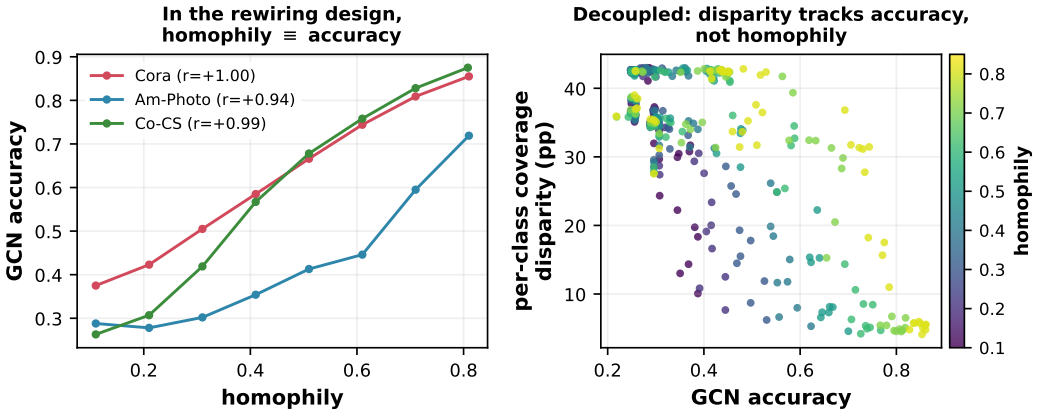

Figure 4: Accuracy, not homophily, drives per-class severity. *Left:* in the rewiring design, homophily and GCN accuracy move together ($r = 0.94$–1.00). *Right:* decoupled via training checkpoints, disparity is a function of accuracy (x-axis); homophily (color) does not separate it.

C.2 Accuracy, not homophily, sets per-class severity

Coverage disparity is largest on heterophilous graphs, so the literature implicitly treats structure as the driver. A controlled experiment points to classifier accuracy instead. We use degree-preserving rewiring to drive homophily from 0.8 to 0.1 while keeping the degree sequence and class sizes exactly fixed (Fig. 3); disparity rises as homophily falls (correlation -0.83 to -0.99), but in this design homophily and GCN accuracy move together ($r = 0.94$–1.00; Fig. 4, left). Pulling them apart with training checkpoints, where accuracy varies within one homophily level, homophily has no effect at matched accuracy: the partial correlation is -0.07 on Amazon-Photo and $+0.44$ on Cora, and heterophilous points show no more disparity than homophilous ones at equal accuracy (28.3 vs 33.4pp on Cora; 41.4 vs 40.1 on Amazon-Photo; descriptive contrasts, as checkpoints within a run are autocorrelated). A less accurate model spreads its errors unevenly across classes; heterophily matters by lowering accuracy. This is the empirical face of Result 1, coverage failures following the model's error-and-confidence profile, which is why the fix is calibration rather than architecture, and it cautions the growing structure-aware conformal literature against conditioning on structural proxies for what is ultimately an accuracy-and-confidence variable.

D Terminology and Abbreviations

Abbreviations used throughout the paper are expanded in Table 5.

Table 4: The paper's technical vocabulary at a glance, with where each term is defined.

Term	Meaning	Defined
Prediction set $C(v)$	candidate labels output for node v	§3
Marginal coverage	$\Pr(y_v \in C(v))$ averaged over *all* test nodes	§3
Conditional coverage	the same probability within a subgroup (here: a stratum)	§3
Nonconformity score	scalar strangeness of a label for a node (LAC/APS/RAPS/DAPS)	§3
Calibration set	held-out labeled nodes that set the threshold $\hat{q}$	§3
$\hat{q}$	the $\lceil (1-\alpha)(n+1) \rceil$-th smallest calibration score	§3
Hop distance $d(v)$	shortest-path distance from v to the nearest training node	§4
Distance stratum	nodes grouped by $d(v) \in \{1, 2, 3, \geq 4\}$	§4
Training-set shadow	systematic under-coverage of the $d{=}1$ stratum	§4
Shadow gap	target coverage minus coverage at $d{=}1$	§4
Message passing	GNN layers averaging neighbor representations	§5
Distance-Mondrian	computing $\hat{q}$ separately within each distance stratum	§7

Table 5: Model, score, and method abbreviations used in the paper (alphabetical), with full forms and the originating reference where applicable.

Abbreviation	Full form	Reference
APPNP	Approximate Personalized Propagation of Neural Predictions	Gasteiger et al. (2019)
APS	Adaptive Prediction Sets	Romano et al. (2020b)
BFS	Breadth-First Search	–
BH	Benjamini–Hochberg (multiple-testing correction)	–
CF-GNN	Conformalized Graph Neural Network	Huang et al. (2023)
ChebNet	Chebyshev spectral graph convolutional network	–
DAPS	Diffusion Adaptive Prediction Sets	Zargarbashi et al. (2023)
GAT	Graph Attention Network	–
GCN	Graph Convolutional Network	–
GNN	Graph Neural Network	–
GNTK	Graph Neural Tangent Kernel	Du et al. (2019)
GraphGPS (GPS)	General, Powerful, Scalable graph transformer	–
GraphSAGE	Graph Sample and Aggregate	–
LAC	Least Ambiguous set-valued Classifier	Sadinle et al. (2019)
MLP	Multilayer Perceptron	–
NAPS	Neighbourhood Adaptive Prediction Sets	Clarkson (2023)
NMI	Normalized Mutual Information	–
NTK	Neural Tangent Kernel	Jacot et al. (2018)
ogbn-arxiv	Open Graph Benchmark arXiv citation network	–
PPR	Personalized PageRank	–
RAPS	Regularized Adaptive Prediction Sets	Angelopoulos et al. (2021)
ReLU	Rectified Linear Unit	–
SGC	Simple Graph Convolution	–
SIGN	Scalable Inception Graph Network	–

E DATASETS

Table 6 lists the 12 node-classification benchmarks: three citation graphs (Planetoid public splits), four co-purchase/co-authorship graphs (deterministic 20-per-class semi-supervised splits), and the five heterophily graphs of Platonov et al. (2023) (public 50% splits, first mask). The first seven are the sparse-label suite where the shadow appears; the last five are the dense-label scope condition. Homophily h is the fraction of edges joining same-class nodes (edge homophily; see Mironov & Prokhorenkova (2024) for a critical treatment of homophily measures). We drop chameleon/squirrel following Platonov et al. (2023).

Table 6: The 12 benchmarks. h = edge homophily; Imb. = largest/smallest class-size ratio.

Dataset	Domain	Nodes	Edges	Feat.	Cls.	h	Imb.
Cora	citation	2,708	5,278	1,433	7	0.81	4.5
CiteSeer	citation	3,327	4,552	3,703	6	0.74	2.7
PubMed	citation	19,717	44,324	500	3	0.80	1.9
Amazon-Photo	co-purchase	7,650	119,081	745	8	0.83	5.9
Amazon-Computers	co-purchase	13,752	245,861	767	10	0.78	17.7
Coauthor-CS	co-author	18,333	81,894	6,805	15	0.81	35.1
Coauthor-Physics	co-author	34,493	247,962	8,415	5	0.93	6.3
Roman-empire	heterophily	22,662	32,927	300	18	0.05	10.1
Amazon-ratings	heterophily	24,492	93,050	300	5	0.38	8.5
Minesweeper	heterophily	10,000	39,402	7	2	0.68	4.0
Tolokers	heterophily	11,758	519,000	10	2	0.60	3.6
Questions	heterophily	48,921	153,540	301	2	0.84	32.5

F Experimental Setup

Models. Two layers, hidden width 64, ReLU, dropout 0.5: GCN and GraphSAGE standard convolutions, GAT with 8 heads (ELU), and a feature-only MLP replacing convolutions with linear layers. Adam, learning rate 0.01, weight decay 5×10^{-4}, 200 epochs. GCN and MLP always share features, splits, hidden width, training recipe, and conformal randomization u; only message passing differs. The depth study uses L stacked GCN convolutions (L=1: a single input-to-class convolution). **Conformal.** Split conformal at $1-\alpha=0.9$; the pool is split 50/50 into calibration/test, reshuffled per seed; threshold index $\lceil(1-\alpha)(n+1)\rceil$. RAPS uses $\lambda=0.05$, $k_{\text{reg}}=1$; APS/RAPS use the randomized form with the same uniform draw u shared across scores and architectures within a seed. Shadow strata are hop distances $\{1,2,3,\geq4\}$ to the nearest training node (BFS), with a ≥ 20-node guard per reported stratum; per-stratum test counts are reported in the results files. Per-class coverage disparity is the standard deviation of per-class coverage (classes with ≥ 20 evaluation nodes). All results average 5 seeds (10 for the score-verdict study). **Distance-Mondrian.** Per-distance-stratum thresholds from calibration nodes in the same stratum (≥ 10 nodes, else global fallback), same randomization as the global method. **Moved-mask arm.** In Control 2 the pool is the full complement of the re-sampled mask (no validation carve-out), a deliberate deviation from the fixed-mask protocol since the original validation set is meaningless for a re-drawn training set. Table 7 extends this arm to ten masks per dataset.

Table 7: Mask-marginal shadow (10 re-sampled training masks per dataset, GCN vs paired MLP; dip = d=1 vs unweighted farther-strata mean, pp). The Table 1 public/deterministic-mask dips are single draws from these distributions; the Amazon graphs' far-strata collapse dominates their resampled arms (both architectures shift positive).

Dataset	GCN dip (mean±sd)	[min, max]	negative	MLP dip	gap vs target
Cora	-6.8 ± 2.5	$[-10.1,-3.5]$	10/10	$+2.3\pm1.8$	4.6
CiteSeer	-15.4 ± 2.9	$[-19.9,-10.0]$	10/10	$+0.7\pm2.4$	12.8
PubMed	-5.4 ± 2.7	$[-10.2,-2.0]$	10/10	$+0.0\pm2.1$	5.9
Amazon-Photo	-2.3 ± 9.9	$[-9.8,+25.3]$	9/10	$+4.1\pm9.3$	2.1
Amazon-Computers	$+6.3\pm5.0$	$[-0.6,+14.3]$	2/10	$+12.5\pm3.7$	2.1
Coauthor-CS	-5.4 ± 1.7	$[-8.8,-2.4]$	10/10	$+2.3\pm1.2$	5.4
Coauthor-Physics	-5.3 ± 1.2	$[-6.8,-2.3]$	10/10	-0.2 ± 2.2	5.2

Matched-coverage diagnostic. Sets an oracle threshold from test-label quantiles so all models realize identical marginal coverage before disparity is measured; used only as a diagnostic and always reported alongside the deployable number. **Significance.** Paired-by-seed t confidence intervals; a verdict flip is significant when two scores have opposite-sign intervals both excluding zero; robustness claims apply Benjamini–Hochberg over the full 1,320-test robustness family (11 of the 12 benchmarks). **Multiplicity across the shadow tables:** applying BH at q=0.05 within each table family, every star in the main shadow table, the architecture and ChebNet tables, the width sweep, and the community ablation survives; 5 of the 105 starred cells (132 tested) lose significance, all

peripheral (APPNP Amazon-Photo $K{=}1$; SIGN Cora and CiteSeer $K{=}10$ and Coauthor-Physics $K{=}5$; NTK Cora), and the affected claims are qualified where they appear. Compute: Apple-Silicon CPU.

G THE SCORE-DEPENDENCE IS NOT GRAPH-SPECIFIC

The score-dependence of §C.1 also appears, in magnitude, on non-graph benchmarks (tabular Forest CoverType, image digits, text 20 Newsgroups; e.g. $+0.8$pp under LAC vs $+2.8$pp under APS on 20 Newsgroups). Graphs are where it amplifies into sign reversals of published verdicts, and where the shadow (§4), a purely graph-positional phenomenon with no tabular analogue, lives.

H PROOF OF RESULT 2

Setting and notation. $X \in \mathbb{R}^{n\times p}$ stacks node features; $S \in \mathbb{R}^{n\times n}$ is the symmetrically normalized adjacency with self-loops, $S = \tilde{D}^{-1/2}(A+I)\tilde{D}^{-1/2}$ where $\tilde{D}$ is the diagonal degree matrix of $A+I$; the L-layer linear GNN (SGC) computes representations $Z = S^L X$, whose row $z_v \in \mathbb{R}^{1\times p}$ is node v's representation. T is the training set; $Z_T \in \mathbb{R}^{|T|\times p}$ and $Y_T \in \mathbb{R}^{|T|\times C}$ collect the training rows of Z and the one-hot label rows $y_u \in \mathbb{R}^{1\times C}$, $u \in T$. The model solves the ridge problem $W^\star = \arg\min_W \|Z_T W - Y_T\|_F^2 + \lambda\|W\|_F^2$, $\lambda > 0$, and predicts $f(v) = z_v W^\star \in \mathbb{R}^{1\times C}$.

Step 1 (closed form of the ridge solution). The objective is strictly convex for $\lambda > 0$ (its Hessian in W is $2(Z_T^\top Z_T + \lambda I) \succ 0$), so the unique minimizer is obtained by setting the gradient $2Z_T^\top(Z_T W - Y_T) + 2\lambda W$ to zero:

$$W^\star \;=\; (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top Y_T,$$

where the inverse exists because $Z_T^\top Z_T + \lambda I \succeq \lambda I \succ 0$.

Step 2 (every prediction is a training-label smoother). For any node v,

$$f(v) \;=\; z_v W^\star \;=\; z_v (Z_T^\top Z_T + \lambda I)^{-1} Z_T^\top Y_T.$$

Write $Z_T^\top Y_T = \sum_{u\in T} z_u^\top y_u$ (the u-th column of $Z_T^\top$ is $z_u^\top$ and the u-th row of Y_T is y_u). Substituting,

$$f(v) \;=\; \sum_{u\in T} \underbrace{\left[z_v (Z_T^\top Z_T + \lambda I)^{-1} z_u^\top\right]}_{=:\, k(v,u)\,\in\,\mathbb{R}} \; y_u,$$

which is the claimed smoother: the prediction for *any* node is a k-weighted combination of training labels, with a data-dependent scalar kernel $k(v,u)$.

Step 3 (the structural self-feature term). By definition of $Z = S^L X$,

$$z_v \;=\; \sum_w S^L(v,w)\, x_w,$$

so the representation of v contains training node u's *own raw features* x_u with scalar weight $S^L(v,u)$. Since S has self-loops and nonnegative entries, $S^L(v,u) > 0$ if and only if there is a walk of length at most L from v to u, i.e. exactly when u is within L hops of v; for all farther u the weight is zero. For the feature-only model ($S = I$): $z_v = x_v$, so for $v \neq u$ the representation of v contains no copy of x_u and the kernel $k(v,u)$ depends on v only through v's own features.

Conclusion. For a training node u within L hops of v, the kernel weight $k(v,u) = z_v(Z_T^\top Z_T + \lambda I)^{-1} z_u^\top$ is evaluated at a z_v that literally contains a positively weighted copy of x_u, the very features the loss was fit to: a structural term absent for the MLP and absent for all nodes beyond L hops. This is the label-bleed asymmetry between the training set's L-hop neighborhood and the rest of the graph claimed in the result. □

Remarks. (i) The argument uses only strict convexity, the rank-one expansion of $Z_T^\top Y_T$, and the support pattern of S^L; no assumption on the features beyond finiteness. (ii) The result is a statement about the *linear* (SGC) model; for finite-width nonlinear architectures we make no formal claim (Result 4 covers the infinite-width regime); the propagation dials (App. K) and the architecture ordering (Table 20) are the evidence that the same asymmetry operates there. (iii) The kernel magnitude, not just its support, matters in practice; the microtest of §6 probes exactly the sign structure ($y_v = y_u$ vs $y_v \neq y_u$) that the smoother form predicts.

I SUPPLEMENTARY FIGURES

Figure 5 gives two node-level views behind §6 and Control 4. *Left:* stratifying test nodes by *local homophily* quartiles, a global threshold under-covers heterophilous-neighborhood (low-confidence) nodes on three of four graphs; Coauthor-CS is flat, and on Amazon-ratings the underlying model is weak (accuracy 0.37), so that panel reflects the low-accuracy regime rather than differential redistribution. *Right:* per-class coverage disparity is essentially flat in GCN depth and separated by graph difficulty, complementing the depth control. Figure 6 shows the class-conditional (Mondrian) repair of per-class disparity across all 12 benchmarks referenced in §7.

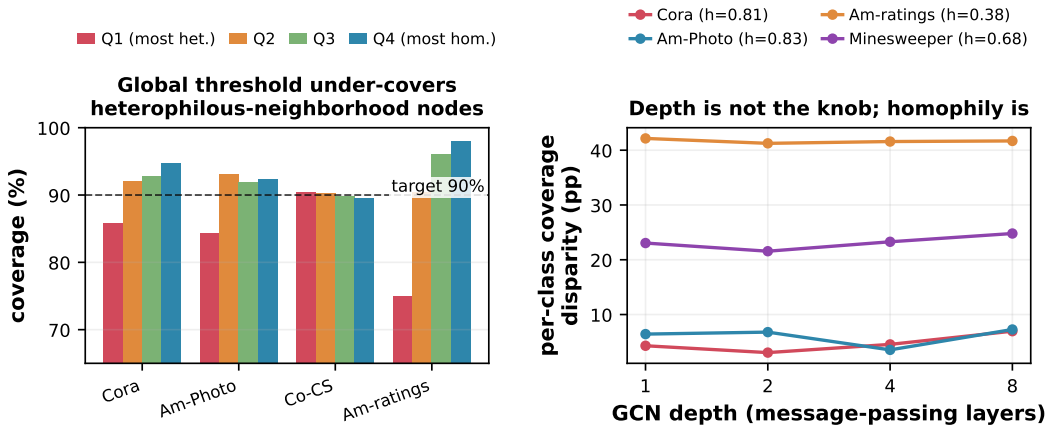

Figure 5: Node-level views. *Left:* coverage by local-homophily quartile (Q1 most heterophilous) under a global threshold. *Right:* per-class coverage disparity vs GCN depth.

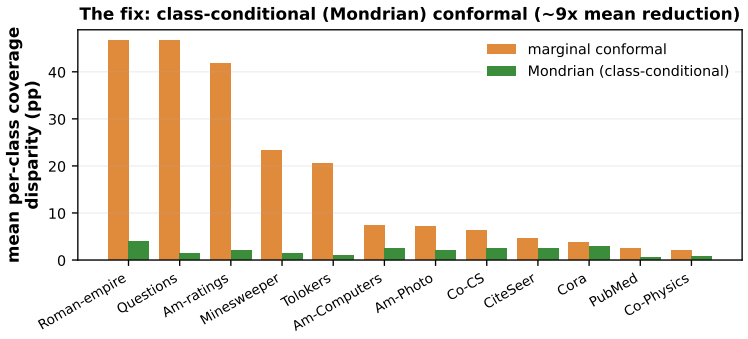

Figure 6: Class-conditional (Mondrian) calibration cuts mean per-class coverage disparity from 17.7 to 2.0pp across the 12 benchmarks (2.4× on the homophilous seven; the larger heterophily factor is partly repair of marginal under-coverage, §7).

J FALLBACK AND CALIBRATION-SPLIT AUDIT

We instrument Algorithm 1's sparse-stratum fallback across calibration/test splits $f \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$ of the pool (5 seeds; 4 strata incl. $d{=}3$ and $d{\geq}4$ separately; unreachable calibration nodes fold into the top stratum, as the algorithm bins them). Table 8: the fallback *never triggers* (0 of 25 seed×ratio cells per dataset), and both the shadow (global $d{=}1$ coverage) and the fix are stable across the entire sweep. Starving the budget to $f{=}0.02$ makes it fire exactly where $n_c < n_{\min}{=}10$ (e.g. PubMed $d{=}1$: $n_c \approx 4.8$, all seeds), and the affected stratum then reverts to the *global* threshold's coverage (PubMed $d{=}1$: 0.756 vs the global method's 0.776): the guarantee degrades to marginal-only rather than failing unpredictably. At $f{=}0.05$ the fallback has already largely disappeared.

Table 8: Fallback audit (5 seeds, α=0.1, $n_{\min}$=10). Mean min. n_c: the smallest per-stratum calibration count at the leanest realistic split (f=0.2), averaged over seeds (per-seed minima run slightly lower). Ranges are over $f \in [0.2, 0.8]$.

Dataset	Mean min. n_c (f=0.2)	Fallbacks	Global d=1 cov.	Mondrian d=1 cov.
Cora	76	0/25	0.881–0.892	0.911–0.918
CiteSeer	58	0/25	0.823–0.846	0.882–0.910
PubMed	53	0/25	0.776–0.799	0.886–0.912
Amazon-Photo	33	0/25	0.851–0.865	0.894–0.910
Amazon-Computers	82	0/25	0.858–0.863	0.897–0.911
Coauthor-CS	321	0/25	0.838–0.850	0.887–0.909
Coauthor-Physics	219	0/25	0.852–0.858	0.892–0.905

K PROPAGATION CONTROL: APPNP

Table 9 reports the paired-by-seed d=1 dip for the fixed-capacity APPNP control of §5 (Control 4): a two-layer MLP (identical parameters for all K) followed by K parameter-free personalized-PageRank steps (teleport 0.1); K=0 reproduces the MLP arm bit-identically. The shadow is off at K=0, on at K=1 (6/7 significant), and saturates for $K \geq 2$, the signature of geometric PPR convergence; Amazon-Computers inverts, consistent with its weak-mixing behavior in Table 20.

Table 9: APPNP: d=1 dip vs farther strata (pp, mean $\pm$ sd, 5 seeds; * significant, paired t). Identical parameterization at every K.

Dataset	K=0 (MLP)	K=1	K=2	K=5	K=10
Cora	-0.2 ± 1.4	$-3.6 \pm 1.9^*$	$-4.7 \pm 2.2^*$	$-3.8 \pm 2.2^*$	$-3.4 \pm 2.4^*$
CiteSeer	$+2.7 \pm 3.2$	$-8.2 \pm 1.6^*$	$-9.5 \pm 2.4^*$	$-9.3 \pm 2.7^*$	$-7.7 \pm 2.4^*$
PubMed	-1.7 ± 3.2	$-9.0 \pm 3.6^*$	$-9.9 \pm 4.0^*$	$-9.6 \pm 1.9^*$	$-9.4 \pm 2.0^*$
Amazon-Photo	$+1.6 \pm 0.8^*$	$-2.0 \pm 1.6^*$	$-3.8 \pm 1.1^*$	$-5.1 \pm 1.1^*$	$-5.0 \pm 1.4^*$
Amazon-Computers	$+6.7 \pm 1.9^*$	$+11.5 \pm 2.6^*$	$+8.6 \pm 2.7^*$	$+6.7 \pm 3.1^*$	$+8.3 \pm 2.6^*$
Coauthor-CS	$+1.7 \pm 1.1^*$	$-5.5 \pm 1.2^*$	$-5.6 \pm 1.3^*$	$-5.3 \pm 1.5^*$	$-4.8 \pm 1.6^*$
Coauthor-Physics	$+0.3 \pm 1.0$	$-5.0 \pm 2.0^*$	$-4.6 \pm 1.8^*$	$-4.2 \pm 2.1^*$	$-3.9 \pm 2.1^*$

Input-side decoupling (SIGN/SGC-style). The complement of the APPNP dial: pre-propagated features $S^K X$ (symmetric-normalized operator, plain powers, no teleport) are fed to the *same* fixed two-layer MLP. This is exactly Result 2's featurization $Z = S^L X$ with the linear head replaced by a nonlinear one: a nonlinear MLP with explicit propagation. Table 10: the shadow is off at K=0 (bit-identical to the MLP control) and on at K=1 on 6/7 datasets: the containment of training-node features in $(S^K X)_v$ survives the nonlinear head, confirming Result 2's mechanism in the finite nonlinear regime from the input side. Because plain powers oversmooth (no teleport), large K degrades accuracy (Amazon-Photo d=1 accuracy 0.77 $\to$ 0.57 by K=10) and the dips in that regime are reported but not interpreted; Amazon-Computers inverts at small K (its family pattern), and its large-K negative dips occur at collapsed accuracy (0.22–0.31), a degenerate regime we flag rather than claim.

L NONLINEAR THEORY: THE NTK REGIME

Result 2 is proven for linear GNNs. Here we show the label-bleed structure is not an artifact of linearity: in the infinite-width limit, a *nonlinear* (ReLU) GCN is provably the same kind of training-label smoother, with the same L-hop self-feature support.

Result 4 (Label bleed survives nonlinearity; NTK regime)**.** *Let F be the two-layer ReLU GCN $F(X) = \frac{\sqrt{c_\sigma}}{\sqrt{h}} S\,\sigma(SXW_1)\,W_2$ (linear output head, propagation S in both layers, NTK parameterization, $c_\sigma = 2$, with mirrored (antisymmetric) initialization so the network's initial function is identically zero), trained by gradient flow on the squared loss over the training rows with penalty $\lambda\|\theta - \theta_0\|^2$ about the initialization, in the infinite-width limit $h \to \infty$. Then: (i) every prediction is*

Table 10: Input-side propagation (SIGN/SGC-style): d=1 dip (pp, mean $\pm$ sd, 5 seeds; * significant). Same fixed MLP at every K; K=0 reproduces the MLP control bit-identically. Daggers: d=1 accuracy below 0.5 (oversmoothed, not interpreted).

Dataset	K=0	K=1	K=2	K=5	K=10
Cora	-0.2 ± 1.4	$-3.5 \pm 2.2^*$	$-5.2 \pm 2.7^*$	$-5.1 \pm 3.2^*$	$-2.1 \pm 1.6^*$
CiteSeer	$+2.7 \pm 3.2$	$-7.1 \pm 2.8^*$	$-8.9 \pm 3.2^*$	$-5.7 \pm 3.9^*$	$-4.3 \pm 3.1^*$
PubMed	-1.7 ± 3.2	$-9.7 \pm 3.3^*$	$-10.8 \pm 3.5^*$	$-11.1 \pm 1.9^*$	$-7.3 \pm 2.8^*$
Amazon-Photo	$+1.6 \pm 0.8^*$	$-2.6 \pm 0.6^*$	$-3.8 \pm 0.6^*$	-0.2 ± 2.3	-0.6 ± 2.0
Amazon-Computers	$+6.7 \pm 1.9^*$	$+7.5 \pm 2.7^*$	$+3.9 \pm 3.3$	$-6.1 \pm 2.5^{*\dagger}$	$-10.3 \pm 1.1^{*\dagger}$
Coauthor-CS	$+1.7 \pm 1.1^*$	$-5.4 \pm 1.2^*$	$-5.6 \pm 1.2^*$	$-4.4 \pm 1.3^*$	$-3.2 \pm 0.8^*$
Coauthor-Physics	$+0.3 \pm 1.0$	$-4.7 \pm 1.8^*$	$-4.6 \pm 2.5^*$	$-3.5 \pm 2.5^*$	-3.2 ± 2.8

a training-label smoother, $F(v) = \sum_{u \in T} k(v,u)\, y_u$ with $k(v,\cdot) = \Theta_{v,T}\,(\Theta_{T,T} + \lambda I)^{-1}$, where Θ is the GCN neural tangent kernel; (ii) $\Theta = S(\Sigma_1 \odot \dot{E}_1)S^\top + SE_1S^\top$ with $\Sigma_1 = SXX^\top S^\top$ and $E_1, \dot{E}_1$ entrywise arc-cosine transforms of Σ_1, so Θ_{vu} is a function only of the aggregated-feature inner products $\{(SX)_a \cdot (SX)_b : a \in \bar{N}(v),\, b \in \bar{N}(u)\}$; (iii) $(SX)_a$ contains x_u with weight $S(a,u) > 0$ exactly when $a \in \bar{N}(u)$; hence for v within L=2 hops of a training node u, Θ_{vu} evaluates inner products of vectors that both contain x_u: the structural self-feature term of Result 2, now inside the nonlinear kernel. For $S = I$ (the MLP-NTK) no such structural asymmetry exists. For APPNP (propagation after the network), $\Theta_{\mathrm{APPNP}} = \Pi_K\, \Theta_{\mathrm{MLP}}\, \Pi_K^\top$ exactly, where Π_K is the K-step personalized-PageRank operator.

Proof. (i): in the infinite-width limit, gradient descent on the squared loss tracks the linearization of the network about its initialization (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Arora et al., 2019), so the limit prediction is $f_0(v) + \Theta_{v,T}(\Theta_{T,T} + \lambda I)^{-1}(Y_T - f_0(T))$, where f_0 is the initial function; mirrored initialization sets $f_0 \equiv 0$, and the penalty $\lambda\|\theta - \theta_0\|^2$ is quadratic in the linearized coordinates, so the function-space solution is kernel ridge regression with the network's NTK and regularization λ, giving the stated smoother. The node-level GCN-NTK and its ReLU arc-cosine form are derived by Sabanayagam et al. (2023) (Thm. 1 and Cor. 2; see also the graph NTK of Du et al., 2019). Linear-output specialization: with an identity output activation the top layer's derivative kernel is identically one, so the output-layer arc-cosine factor in their recursion is the all-ones matrix; substituting into their Cor. 2 leaves exactly the two displayed terms. (ii) By their recursion, E_1 and $\dot{E}_1$ are entrywise functions of Σ_1 and its diagonal; pre- and post-multiplication by S extends the dependence of entry (v,u) exactly to the closed one-hop neighborhoods $\bar{N}(v), \bar{N}(u)$, whose Σ_1 entries are the stated inner products. (iii) $(SX)_a = \sum_w S(a,w)x_w$ contains x_u iff $S(a,u) \neq 0$, i.e. iff $a \in \bar{N}(u)$ (self-loops included); for $S = I$, $(X)_a = x_a$ never contains x_u for $a \neq u$. The APPNP identity holds because a fixed linear map applied after the network multiplies the Jacobian on the left, so the kernel transforms covariantly. □

Scope. The NTK regime is lazy training: it shows nonlinearity per se does not destroy the mechanism, not that finite training dynamics are fully captured; the experiments below locate where each regime governs.

Exact-kernel experiment (zero training). We compute the kernels in closed form and run kernel ridge ($\lambda = 10^{-4}\,\overline{\mathrm{diag}}\,\Theta_{TT}$) through the identical conformal pipeline (score $1-f_c$; same splits/seeds; the paired t here measures split variability only, as the kernels are deterministic). Table 11: on the citation graphs the GCN-NTK reproduces the shadow with *no training at all*, the S=I control erases it, and the APPNP-NTK family shows the kernel-regime dose-response in K. On Amazon-Photo the exact kernel shows no d=1 dip (miscoverage sits at d=3) although trained networks dip at every width (Table 12): on that graph the shadow is driven by *feature learning*, which locates where the lazy-regime account ends, consistent with the Amazon family being the deviant throughout (Table 20). Distance-Mondrian restores ≈ 0.90 per stratum for every kernel arm.

Table 11: Exact-kernel shadow (d=1 dip vs farther strata, pp; 5 split-seeds; * paired t, $p < 0.05$). No training: kernel ridge with closed-form NTKs.

Dataset	GCN-NTK	MLP-NTK (S=I)	APPNP-NTK K=1	K=2	K=10
Cora	$-2.8 \pm 2.3^*$	-0.4 ± 1.3	-2.0 ± 2.4	-2.7 ± 2.8	$-3.6 \pm 2.6^*$
CiteSeer	$-7.2 \pm 2.8^*$	$+2.8 \pm 1.8^*$	$-7.5 \pm 4.0^*$	$-6.3 \pm 3.2^*$	$-6.5 \pm 4.1^*$
Amazon-Photo	$+11.2 \pm 1.3^*$	$+2.9 \pm 1.2^*$	$+7.2 \pm 0.8^*$	$+10.2 \pm 1.7^*$	$+12.1 \pm 1.0^*$

Table 12: Width sweep (finite trained GCN, APS; d=1 dip, pp; 5 seeds; W=64 is the paper's model, reproduced bit-identically). The dip is not a finite-width artifact.

Dataset	W=16	W=64	W=256	W=1024
Cora	$-7.0 \pm 3.5^*$	$-4.7 \pm 2.5^*$	$-5.0 \pm 3.2^*$	$-4.4 \pm 1.7^*$
CiteSeer	$-7.9 \pm 4.2^*$	$-9.0 \pm 2.5^*$	$-9.9 \pm 3.0^*$	$-9.2 \pm 2.3^*$
Amazon-Photo	-2.1 ± 3.3	$-8.4 \pm 1.1^*$	$-9.2 \pm 1.3^*$	$-8.8 \pm 1.6^*$

M SET-SIZE PARETO AND CONFIDENCE-BIN ANALYSIS

Table 13: for every score, distance-Mondrian removes almost all of the worst-stratum coverage violation at unchanged mean set size: the d=1 growth reported in §7 is paid for by smaller sets in the over-covered far strata. DAPS-global has the single worst dataset violation ($+0.176$) despite being structure-aware; LAC + distance-Mondrian sits on the Pareto frontier. Table 14: binning test nodes by top-1 probability, coverage under the global threshold is nominal in the moderate bins and decays smoothly from $\hat{p}_{\text{top}} \approx 0.6$ while accuracy rises; d=1 nodes concentrate in the high-confidence bins (cf. Result 1).

Table 13: Pareto: mean set size vs worst per-stratum violation of the 0.9 target (macro-average over 7 datasets, 5 seeds; "worst ds" = single worst dataset; violations computed on seed-averaged per-stratum coverage).

Score	Threshold	Mean size	Max violation (avg)	(worst ds)
APS	global	2.08	$+0.072$	$+0.157$
APS	dist-Mondrian	2.04	$+0.001$	$+0.004$
LAC	global	1.67	$+0.054$	$+0.146$
LAC	dist-Mondrian	1.66	$+0.008$	$+0.013$
RAPS	global	1.92	$+0.059$	$+0.157$
RAPS	dist-Mondrian	1.89	$+0.003$	$+0.009$
DAPS	global	1.66	$+0.047$	$+0.176$
DAPS	dist-Mondrian	1.67	$+0.007$	$+0.026$

N TUNED-TRAINING ARM

To test whether the shadow depends on the fixed training recipe, we re-run the Table 1 protocol with per-(dataset, seed) model selection: a grid over learning rate $\{0.01, 0.005\}$, weight decay $\{5\times10^{-4}, 0\}$, dropout $\{0.5, 0.2\}$, and width $\{64, 256\}$, each trained up to 500 epochs with patience 50 on validation accuracy (best-validation weights restored), the best configuration per seed selected on validation accuracy alone. Selection touches only the train and validation sets, so pool exchangeability is untouched. Table 15: the dip survives at equal or larger magnitude on the three citation graphs and Coauthor-CS, and attenuates to non-significance on Amazon-Photo and Coauthor-Physics. On ogbn-arxiv (2% stratified labels, standard tuned recipe: 3-layer width-256 GCN with BatchNorm, lr 0.01, early stopping on a 5,000-node validation set carved from the pool before the calibration/test split), the tuned model reaches similar accuracy (66 vs 67%) but a different coverage profile: d=1 sits at 91.3–91.9% and the deficit moves entirely to the far strata (78–80% at $d\geq4$), so at that scale the training regime relocates the conditional failure rather than removing it.

Table 14: Coverage by predicted top-1 probability (APS, global threshold; macro-average over datasets meeting the $n \geq 20$ per-bin guard, 5 seeds).

$\hat{p}_{\text{top}}$ bin	n/dataset	Accuracy	Coverage	Share of d=1
$[0.00, 0.40)$	$\sim$1676	0.562	0.924	0.10
$[0.40, 0.50)$	$\sim$562	0.694	0.917	0.19
$[0.50, 0.60)$	$\sim$627	0.793	0.919	0.24
$[0.60, 0.70)$	$\sim$662	0.843	0.872	0.28
$[0.70, 0.80)$	$\sim$907	0.925	0.880	0.22
$[0.80, 0.90)$	$\sim$1234	0.948	0.825	0.31
$[0.90, 0.95)$	$\sim$990	0.952	0.804	0.30
$[0.95, 1.00]$	$\sim$1201	0.986	0.779	0.22

Table 15: Tuned-training arm (5 seeds; dip = d=1 vs unweighted farther-strata mean, pp; * significant, paired t). Fixed-recipe dips from Table 1 for comparison.

Dataset	fixed dip	tuned dip	tuned t	tuned test acc.
Cora	-4.7^*	$-6.5 \pm 3.7^*$	-3.9	0.818
CiteSeer	-9.0^*	$-8.0 \pm 4.2^*$	-4.2	0.692
PubMed	-11.8^*	$-11.8 \pm 3.1^*$	-8.6	0.777
Amazon-Photo	-8.4^*	-2.0 ± 2.1	-2.1	0.894
Amazon-Computers	$+0.1$	-2.0 ± 2.9	-1.6	0.794
Coauthor-CS	-6.0^*	$-7.4 \pm 0.6^*$	-28.2	0.910
Coauthor-Physics	-5.0^*	-2.6 ± 2.2	-2.6	0.931

O DECOMPOSING THE APS SHADOW

Randomized APS can produce empty prediction sets on peaked-correct nodes (Romano et al., 2020b; Angelopoulos & Bates, 2023), so part of an APS coverage dip can in principle be avoided by construction. We re-run the Table 1 protocol (same models, splits, and randomization) under three additional arms: randomized APS with empty sets disallowed (a node whose set would be empty receives its top-1 label, so it is covered exactly when the prediction is correct), deterministic APS ($s = \text{mass above} + p_y$), and LAC. Table 16: the randomized arm reproduces Table 1's dips to rounding; disallowing empty sets leaves a significant confident-error core on CiteSeer, PubMed, and Amazon-Photo while the Coauthor shadows collapse to near zero, matching the microtest's attribution; and deterministic APS widens every disparity, since a conservative global threshold pushes far strata toward ceiling coverage while d=1 confident errors persist. Empty sets concentrate at d=1 on every dataset. Distance-Mondrian applies per arm unchanged (Result 3 is score-agnostic).

Table 16: APS decomposition (5 seeds; dip = d=1 vs unweighted farther-strata mean, pp; * significant, paired t, in either direction). "no-empty" disallows empty sets; "det." is deterministic APS. Empty-set fractions are for randomized APS.

Dataset	APS (rand.)	no-empty	det.	LAC	empty d=1	empty all
Cora	-4.7^*	-0.8	-12.8^*	-0.4	6.2%	3.5%
CiteSeer	-9.0^*	-6.2^*	-22.7^*	-6.6^*	3.3%	0.9%
PubMed	-11.8^*	-8.8^*	-19.6^*	-11.4^*	5.6%	1.8%
Amazon-Photo	-8.4^*	-3.9^*	-12.9^*	-0.6	5.0%	2.2%
Amazon-Computers	$+0.0$	$+0.4$	-1.0	$+2.5$	0.5%	0.2%
Coauthor-CS	-6.0^*	-1.0^*	-11.2^*	$+2.3^*$	12.7%	7.5%
Coauthor-Physics	-5.0^*	-0.4	-15.3^*	$+5.4^*$	12.5%	8.2%

P THE INDUCTIVE SETTING

We remove all edges incident to test nodes during training (unseen-nodes inductive setting), reveal the full graph at inference, and score calibration and test nodes identically; splits are paired with

the transductive arm. Note calibration nodes' *features* were present in the training graph while test nodes' were not, so calibration and test are not exchangeable by construction; this is the deployment reality of the inductive regime. The numbers below are therefore purely empirical, with no coverage guarantee attached; a guarantee-bearing treatment would go through conformal prediction beyond exchangeability (Barber et al., 2023) or the inductive graph constructions of Zargarbashi & Bojchevski (2024), which we leave open. Table 17: the shadow persists ($4/4$ negative, $3/4$ significant), at near-transductive magnitude on the denser graphs and attenuated on the small citation graphs (a test node absent from the training graph receives the concentrated near-train confidence only secondhand, through its neighbors' representations); marginal coverage does not collapse; and distance-Mondrian still restores every $d{=}1$ stratum to 0.90–0.94, since it recalibrates within strata at deployment time.

Table 17: Inductive shadow (5 seeds): $d{=}1$ dip (pp; * significant), transductive dip for comparison, marginal test coverage, and the $d{=}1$ stratum after distance-Mondrian.

Dataset	Inductive dip	Transd. dip	Marginal	Mondrian $d{=}1$
Cora	-2.5 ± 2.4	-4.7	0.923	0.927
CiteSeer	$-3.8 \pm 1.9^*$	-9.0	0.921	0.938
PubMed	$-11.4 \pm 4.0^*$	-11.8	0.900	0.914
Amazon-Photo	$-7.5 \pm 0.8^*$	-8.4	0.905	0.900

Q LOCALIZED BASELINE AND GRAPH TRANSFORMER

NAPS. We run the neighborhood-adaptive prediction sets of Clarkson (2023) (per-test-node quantile from calibration nodes within k hops, uniform weights; the conservative full-set fallback applies when the local order statistic does not exist; NAPS targets inductive delivery, and we evaluate it transductively as the localized-calibration representative). At $k{=}3$ (Table 18), the $d{=}1$ stratum remains the *minimum-coverage* stratum on $7/7$ datasets, while far strata are over-covered up to 0.95–0.99 at mean set sizes 1.35–4.07, i.e. 1.1–1.7$\times$ distance-Mondrian's per dataset; at $k{=}2$ the conservatism is severe (mean size up to 5.1; on CiteSeer 76% of test nodes receive the full label set). Localizing by *neighborhood* neither equalizes distance strata nor approaches the Pareto frontier of Table 13: the conditioning axis is what matters.

Table 18: NAPS ($k{=}3$, APS, 5 seeds): coverage by distance stratum, mean set size, fraction of full-label-set (infinite) predictions, and median local calibration count.

Dataset	Cov. $d{=}1$ / 2 / 3 / ≥ 4	Mean size	Full-set frac.	Med. n_{loc}
Cora	0.918 / 0.943 / 0.933 / 0.981	3.12	0.170	32
CiteSeer	0.922 / 0.942 / 0.965 / 0.992	4.07	0.433	11
PubMed	0.842 / 0.912 / 0.924 / 0.950	1.78	0.056	95
Amazon-Photo	0.863 / 0.943 / 0.961 / –	2.12	0.001	1354
Amazon-Computers	0.860 / 0.935 / 0.958 / 0.950	3.40	0.002	4087
Coauthor-CS	0.865 / 0.923 / 0.948 / 0.960	2.53	0.025	310
Coauthor-Physics	0.884 / 0.902 / 0.929 / 0.951	1.35	0.010	665

Community-Mondrian ablation (the conditioning axis, isolated). Community-based Mondrian variants (Zhang et al., 2025) give exact cell-conditional coverage on *communities*; we ablate that conditioning variable with our own machinery: Mondrian with per-cell quantiles (same $n_{\min}{=}10$ fallback) over Louvain communities (structure-only, deterministic; 23–471 communities per graph). Result (Table 19): the partition is valid on its own axis (seed-averaged per-community coverage spans 0.81–1.00, with per-seed minima reaching 0.77, consistent with binomial spread around the 0.9 target). Fallback disclosure: communities with fewer than $n_{\min}{=}10$ calibration nodes revert to the global threshold, which affects most of the many tiny Louvain communities on four datasets (Cora $82/102$, CiteSeer $448/471$, Amazon-Photo $136/151$, Amazon-Computers $315/331$, seed means) but few on the rest (PubMed $5/40$, Coauthor-CS $7/26$, Coauthor-Physics $4/23$); because Louvain communities are size-skewed, the large communities that hold most nodes do receive their own thresholds on every dataset. To remove the concern entirely, we re-ran the ablation on a coarser

partition that deterministically merges every community with fewer than 40 pool nodes into its most-connected neighbor (the merge rule depends only on the graph and the train/val masks, so exactness is preserved): every cell *containing pool nodes* holds ≥ 57 of them (cells whose members are all train/val nodes survive unmerged but hold no calibration or test node, and the recorded fallbacks fire only on those empty cells), so per-community thresholds are active for every evaluated node, and the d=1 dip persists on $7/7$ datasets ($6/7$ significant; Cora -4.1 ± 1.9, CiteSeer -8.9 ± 2.8, PubMed -10.3 ± 4.0, Amazon-Photo -7.5 ± 1.5, Coauthor-CS -6.1 ± 1.7, Coauthor-Physics -4.3 ± 1.9; Amazon-Computers -1.4 ± 3.1 n.s., its family null), close to the global threshold's magnitude and now with no fallback caveat, while distance-Mondrian eliminates it. Training-adjacent nodes occur inside every community, so community-conditional validity neither detects nor removes the shadow: what matters is not Mondrian machinery but the conditioning *axis*.

Table 19: Conditioning-axis ablation: d=1 dip (pp, 5 seeds; * significant) under the global threshold, community-Mondrian (Louvain), and the d=1 coverage under distance-Mondrian (Table 1). Exact validity on the wrong axis leaves the shadow intact.

Dataset	Global dip	Community-Mondrian dip	dist-Mondrian d=1 cov.
Cora	$-4.7 \pm 2.5^*$	$-5.0 \pm 2.3^*$	0.911
CiteSeer	$-9.0 \pm 2.5^*$	$-8.4 \pm 2.1^*$	0.905
PubMed	$-11.8 \pm 3.9^*$	$-10.4 \pm 4.0^*$	0.911
Amazon-Photo	$-8.4 \pm 1.1^*$	$-7.7 \pm 1.5^*$	0.901
Amazon-Computers	$+0.1 \pm 2.5$	-0.9 ± 3.1	0.905
Coauthor-CS	$-6.0 \pm 1.7^*$	$-6.3 \pm 1.7^*$	0.904
Coauthor-Physics	$-5.0 \pm 2.2^*$	$-4.3 \pm 1.9^*$	0.897

GraphGPS (preliminary). A small GraphGPS (two GPS layers: local GCN + global multihead attention, random-walk structural encodings appended to features) run on the three graphs where exact quadratic attention is feasible for us, with the shared training recipe *untuned* for the architecture; its accuracy is accordingly well below the GCN's, so we report this as preliminary. The distance alignment weakens on $2/3$ datasets: on Cora the d=1 dip essentially vanishes (-0.8 ± 2.4, n.s.), and on Amazon-Photo the worst-covered stratum moves *off* d=1 entirely (d=3 at 0.834; dip $+1.1 \pm 4.9$, n.s.), so global attention appears to *redistribute* conditional miscoverage rather than eliminate it, while CiteSeer retains a strong dip (-8.5 ± 3.4, significant). Distance-Mondrian restores ≈ 0.90 per stratum on the GPS scores as well: the fix is model-agnostic. A properly tuned transformer study is future work; we do not evaluate k-WL-style higher-order networks, whose additional power targets isomorphism expressivity rather than confidence concentration, while propagation *order* is already probed by ChebNet (Table 20) and the APPNP sweep (Table 9).

R Additional robustness

Table 20: The shadow across architectures: paired-by-seed d=1 dip vs farther strata (pp, mean $\pm$ sd over 5 seeds; * significant at $p < 0.05$, $t_{\text{crit}} = 2.78$). Negative = under-coverage of the training set's neighbors. The mechanism predicts, before looking, that the dip should track neighbor-mixing strength: GraphSAGE's root weight and GAT's attention dilute neighbor mixing, weakening the shadow, and the observed ordering matches the prediction; Amazon-Computers' farther-strata mean is confounded by its far-strata collapse for all architectures (§4). ChebNet: order-3 Chebyshev filters, 2-hop receptive field per layer.

Dataset	GCN	GAT	GraphSAGE	ChebNet
Cora	$-4.7 \pm 2.5^*$	$-5.2 \pm 3.3^*$	-4.0 ± 3.4	$-4.5 \pm 2.6^*$
CiteSeer	$-9.0 \pm 2.5^*$	$-8.9 \pm 3.4^*$	$-6.2 \pm 3.4^*$	$-6.5 \pm 3.5^*$
PubMed	$-11.8 \pm 3.9^*$	$-8.0 \pm 3.7^*$	$-10.1 \pm 3.5^*$	$-6.4 \pm 4.3^*$
Amazon-Photo	$-8.4 \pm 1.1^*$	$+0.7 \pm 0.9$	$+0.2 \pm 1.4$	-2.5 ± 2.2
Amazon-Computers	$+0.1 \pm 2.5$	$+2.1 \pm 3.3$	$+9.7 \pm 2.4^*$	$+0.1 \pm 2.4$
Coauthor-CS	$-6.0 \pm 1.7^*$	$-3.8 \pm 1.1^*$	-0.3 ± 1.2	$-4.9 \pm 1.5^*$
Coauthor-Physics	$-5.0 \pm 2.2^*$	-1.6 ± 1.4	-2.0 ± 2.3	$-4.3 \pm 2.2^*$

The 11-benchmark score-dependence survives GAT and GraphSAGE backbones, DAPS, three disparity metrics (standard deviation, max−min spread, worst-class undercoverage gap), and Benjamini–Hochberg correction over the full 1,320-test robustness family (11 of the 12 benchmarks). On three datasets with genuine demographic protected attributes (NIFTY Credit, German, Bail; protected attribute excluded from features), no score yields a significant group-coverage reversal ($0/3$); indeed these datasets have near-zero group-label alignment (normalized mutual information ≈ 0.002–0.005), consistent with the instability being class-structured and reaching groups only through label alignment. A score-free probability-integral-transform calibration diagnostic we explored measures an orthogonal axis (calibration rather than coverage) and does not predict the coverage flips.