SEPARATING INFORMATION LIMITS FROM REALIZATION GAPS IN LOCAL GRAPH DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph distillation reports one teacher–student discrepancy, mixing unavailable teacher-relevant information with excess relative to the best predictor under the declared observation. We separate them using $Q_k = \mathbb{E}[P \mid A, Z_k]$, where A declares training-side information and Z_k is the radius-k observation. This standard expected-KL decomposition yields the information term Δ_k; our contribution is to define and diagnose it under an explicit information set and compatible-world law. The observable pair (A, Z_k) does not determine that law, and one complete realization cannot uniformly recover a law-dependent Δ_k without additional assumptions. We assign exact values to newly observed shells and characterize the information boundary through predictive Fisher curvature and conditional Green covariance. On Cora, PubMed, and Roman-empire, matched audits estimate all KL components under post-encoder graph-correlated completion laws. Conditioning on realized training-side logits changes radius-zero information shares from 12.5%, 17.9%, and 80.9% to 2.3%, 1.7%, and 23.7%, reversing the Roman-empire diagnosis. Predictive checks assess the adequacy of these model-conditional values, while qualitative conditioning-set diagnoses remain stable across retained laws. A contractive nonlinear teacher further tests the accounting beyond linear diffusion. The diagnosis is therefore relative to the declared information set and law, not intrinsic to one graph.

1 INTRODUCTION

A global graph teacher can use nodes and edges that a deployable local student cannot observe. Distillation methods improve transfer through response matching, reliability weighting, positional signals, frequency separation, and intermediate supervision (Hinton et al., 2015; Zhang et al., 2022; Wu et al., 2023b; Tian et al., 2023; Wu et al., 2023a; Yang et al., 2024; Zhou et al., 2025; Wu et al., 2024b; Lu et al., 2024; Eskandari et al., 2026). Evaluations typically report the teacher–student discrepancy after training. That number combines missing teacher-relevant information under a declared observation with finite-student excess relative to its conditional oracle.

This distinction matters for both architecture and evaluation. Increasing capacity can reduce finite-student excess but cannot reconstruct an unobserved remote variable. Increasing graph radius can reveal useful information, but a finite learner may still use the larger input less effectively. A performance curve alone therefore cannot determine whether the remedy is more context, a stronger learner, or a different training protocol.

Our answer is an information diagnostic framework. Let P be the frozen teacher distribution, let A denote a declared training-side sigma-field, and let Z_k be the student's nested radius-k observation. Here A defines the estimand, not necessarily the learner's literal optimization transcript. Hence the oracle-relative realization gap is not optimization error alone. The optimal conditional student is $Q_k = \mathbb{E}[P \mid A, Z_k]$. Every measurable local student S_k satisfies

$$\mathbb{E}D_{\mathrm{KL}}(P\|S_k) = \underbrace{\mathbb{E}D_{\mathrm{KL}}(P\|Q_k)}_{\Delta_k\text{: information limit}} + \underbrace{\mathbb{E}D_{\mathrm{KL}}(Q_k\|S_k)}_{\mathcal{E}_k\text{: realization gap}}. \tag{1}$$

Equation (1) is a standard conditional KL projection, not a new information identity. Our contribution is to turn this projection into a graph-distillation diagnostic framework: we define the estimand under

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.

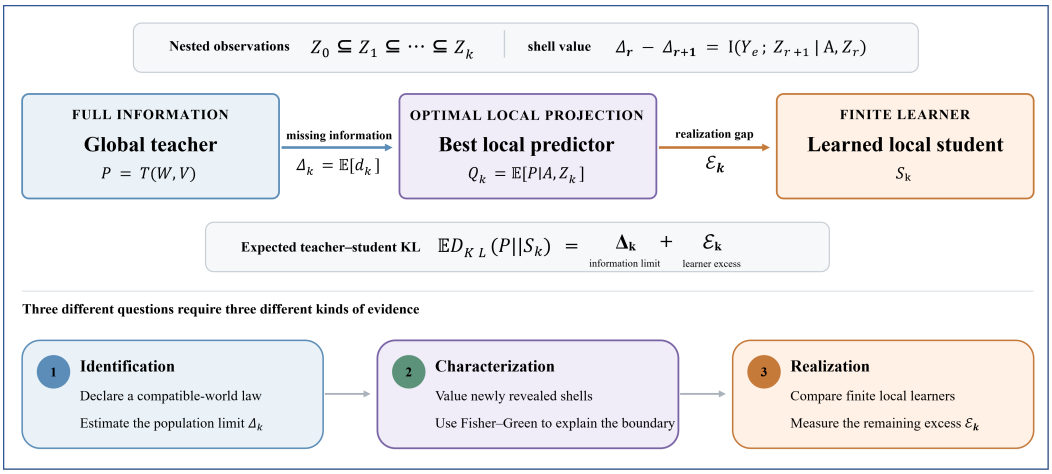


Figure 1: **Information diagnostic framework.** The expected teacher–student KL decomposes into an information limit and a learner-dependent realization gap under the declared population. Nested graph observations assign an information value to each additional shell. Identification, characterization, and realization require different evidence.

an explicit information set and compatible-world law, characterize its graph-specific information boundary, and audit the remaining finite-student excess.

Our contribution follows three stages. **Identification** defines the limit under explicit A, Z_k, and a compatible-world law, separating observable non-identification from failure of uniform recovery from one graph. **Characterization** uses exact shell values and finite-scale Fisher–Green analysis to locate teacher-relevant information. **Realization** uses matched audits to separate the population limit from oracle-relative excess across conditioning sets, laws, and capacities. Our target is fidelity to a fixed teacher; Δ_k represents missing task-label information only when the teacher is the Bayes posterior.

2 Related Work

2.1 GNN-to-MLP and graph distillation

Graph neural networks aggregate neighborhood information (Gilmer et al., 2017; Kipf & Welling, 2017; Hamilton et al., 2017). GLNN distills a GNN into a graphless student (Zhang et al., 2022). KRD, FF-G2M, NOSMOG, VQGraph, and TINED add reliability sampling, frequency-aware targets, positional and adversarial signals, a structure-aware codebook, and layerwise matching (Wu et al., 2023b;a; Tian et al., 2023; Yang et al., 2024; Zhou et al., 2025). These methods primarily ask how to transfer more graph knowledge. HGMD changes training supervision through hardness-aware subgraphs, AdaGMLP changes the learner class through an ensemble, and InfGraND supplies the deployed MLP with precomputed multi-hop features (Wu et al., 2024b; Lu et al., 2024; Eskandari et al., 2026). TGS removes the fixed teacher altogether (Wu et al., 2024a). In our notation these interventions change, respectively, training information A, the realization class, the inference information Z_k, or the estimand itself. We instead keep each coordinate explicit and ask how much is transferable under a fixed information set.

2.2 Distillation under asymmetric information

Knowledge distillation permits a teacher to expose information unavailable to the final predictor (Hinton et al., 2015). Learning using privileged information and generalized distillation formalize related train–test asymmetries (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). Graph locality adds a nested spatial structure to this asymmetry. It permits a radius-dependent information profile, an exact value for each newly revealed shell, and a law-aware diagnostic radius. It also makes the training side information A explicit, so test-node teacher targets and positional codes cannot be treated as innocuous implementation details.

2.3 Information projection, geometry, and graph locality

Proper log loss, conditional projection, conditional mutual information, and Fisher local geometry are established ingredients (Banerjee et al., 2005; Cover & Thomas, 2006; Seyfried et al., 2025; Nielsen, 2024). Our graph-specific use of the explicit $(A, Z_k, \mathcal{L})$ estimand exposes the identification boundary, while conditional Green covariance links graph influence, unresolved uncertainty, and predictive curvature. Strong data-processing inequalities provide channel-specific information contraction (Polyanskiy & Wu, 2017), while recent counterexamples caution against interchanging different entropy-contraction coefficients (Caputo et al., 2024). We therefore separate exact shell values from mechanism-specific envelopes. PPNP/APPNP provides an exact global diffusion operator (Klicpera et al., 2019), and contractive implicit GNNs motivate our nonlinear certificate (Baker et al., 2023). Yu et al. analyze learning dynamics and implicit propagation in finite GNN-to-MLP transfer (Yu et al., 2026); we instead fix $\sigma(A, Z_k)$ and characterize the optimal measurable predictor and its identification. Ye et al. maximize conditional mutual information to improve teacher-side conditional-distribution estimation (Ye et al., 2024); here $I(\widetilde{Y}; Z_{k+1} \mid A, Z_k) = \Delta_k - \Delta_{k+1}$ values a newly revealed graph shell. The resulting information diagnostic distinguishes teacher-task locality from message-passing locality (Alon & Yahav, 2021; Di Giovanni et al., 2023).

3 Information Limits of Local Graph Distillation

The projection identity supplies the accounting foundation; the identification boundary and graph-specific characterization provide the substantive theory.

3.1 Local information and the optimal conditional student

Let W be a random complete graph world, V a root, and $P = T(W, V) \in \Delta^{C-1}$ the output of a trained, frozen teacher. The nested observation $Z_k = O_k(W, V)$ contains the radius-k information available at inference, including boundary metadata needed by the propagation rule. The variable A is a declared post-training side-information sigma-field. It includes every random quantity retained and exposed to the diagnostic under the chosen estimand. Quantities declared as fixed design variables are not random coordinates of that population. Radius denotes the declared propagation observation, including boundary normalization metadata, rather than an induced-subgraph-only view. The projection theorem is valid for any declared A, but empirical claims must use the same A in both the compatible-world population and the conditional computation. Write $\mathcal{G}_k = \sigma(A, Z_k)$ and define

$$Q_k = \mathbb{E}[P \mid \mathcal{G}_k], \qquad d_k = \mathbb{E}[D_{\mathrm{KL}}(P \| Q_k) \mid \mathcal{G}_k], \qquad \Delta_k = \mathbb{E}[d_k]. \tag{2}$$

The conditional residual d_k is $\mathcal{G}_k$-measurable and varies with the realized local context. Its expectation Δ_k is the population information limit. Neither quantity is a particular network or training algorithm.

3.2 Information limit versus realization gap

Theorem 1 (Information accounting under log loss)**.** *For every $\mathcal{G}_k$-measurable probabilistic student S_k, whenever the expectations are finite,*

$$\mathbb{E}D_{\mathrm{KL}}(P \| S_k) = \Delta_k + \mathbb{E}D_{\mathrm{KL}}(Q_k \| S_k). \tag{3}$$

The first term is irreducible under the declared information set. The second is excess KL relative to its conditional oracle; it can include approximation, estimation, optimization, or a mismatch between declared side information and the learner's literal inputs. A stronger learner can reduce a removable part without changing Δ_k. Enlarging A or Z_k can change the information limit itself.

Introduce $\widetilde{Y}$ with $\Pr(\widetilde{Y} = c \mid W, V, A) = P_c$. Then

$$\Delta_k = I(\widetilde{Y}; W, V \mid A, Z_k), \qquad \Delta_k - \Delta_{k+1} = \mathbb{E}D_{\mathrm{KL}}(Q_{k+1} \| Q_k) \geq 0. \tag{4}$$

Available information therefore improves the population optimum monotonically. More precisely, the conditional residual satisfies

$$d_k = \mathbb{E}[d_{k+1} + D_{\mathrm{KL}}(Q_{k+1} \| Q_k) \mid \mathcal{G}_k]\,. \tag{5}$$

Thus (d_k) is a nonnegative supermartingale. This conditional-expectation ordering does not imply that $d_{k+1} \leq d_k$ along every realized context path. No monotonicity conclusion follows for a finite learned student either.

3.3 IDENTIFICATION RELATIVE TO THE OBSERVED INFORMATION

Proposition 1 (Non-identification under local observation). *Suppose two worlds share (A, Z_k) but yield different teacher predictions. There exist completion laws inducing the same distribution of the observable (A, Z_k) and different conditional projection gaps. Moreover, a single complete realization cannot guarantee exact recovery of the gap uniformly over completion laws assigning it positive probability.*

A point mass at w_0 yields zero conditional gap. A mixture of w_0 and w_1 yields a positive weighted Jensen–Shannon divergence. These laws agree on the distribution of (A, Z_k), not on that of the complete world; both nevertheless assign positive probability to w_0. Consequently, a real-graph study must declare a compatible-world law, obtain independent graph worlds, or restrict its conclusion to a deterministic certificate. This distinction determines the empirical design in Section 5. Thus a realized graph can inform Δ_k, but cannot determine this law-dependent functional from the observed information alone.

4 CHARACTERIZING THE INFORMATION BOUNDARY

Shell values follow exactly from nested information; Fisher–Green instead gives a finite-scale local characterization of unresolved graph information.

4.1 SHELL INFORMATION AND LAW-AWARE DIAGNOSTIC RADII

Theorem 2 (Shell values and adaptive stopping). *For nested $\mathcal{G}_k$, every refinement from radius k to $j > k$ satisfies*

$$\Delta_k - \Delta_j = \sum_{r=k}^{j-1} \mathbb{E} D_{\mathrm{KL}}(Q_{r+1} \| Q_r) = \sum_{r=k}^{j-1} I(\widetilde{Y}; Z_{r+1} \mid A, Z_r). \tag{6}$$

If $\bigvee_r \mathcal{G}_r$ reveals P, then Δ_k equals the sum of all future shell values. For any bounded stopping time $\tau \in \{0, \dots, K\}$ adapted to $(\mathcal{G}_k)$,

$$\mathbb{E}[d_\tau] = \mathbb{E} D_{\mathrm{KL}}(P \| Q_\tau) = \Delta_0 - \sum_{r=0}^{K-1} \mathbb{E}[\mathbf{1}\{\tau > r\} D_{\mathrm{KL}}(Q_{r+1} \| Q_r)]\,. \tag{7}$$

Each additional shell is therefore an information resource with an exact expected log-loss value. The two radii serve different decision roles:

$$k^*_{\mathrm{pop}}(\varepsilon) = \min\{k : \Delta_k \le \varepsilon\}, \qquad \tau_{\varepsilon,K} = \inf\{k \in \{0, \dots, K\} : d_k \le \varepsilon\} \wedge K. \tag{8}$$

The first is a population design radius. The second is a bounded stopping time because d_k is $\mathcal{G}_k$-measurable, so Eq. (7) applies with $\tau = \tau_{\varepsilon,K}$. The latter can vary across roots and contexts even when a deployed model uses one fixed maximum depth. The event that the tolerance is not reached by K must be reported separately. This is an information-theoretic, law-aware rule: evaluating d_k requires either the declared completion law or a calibrated estimator together with conditional computation from the observed context. Learning a deployable estimator from local evidence alone is outside our scope.

4.2 PROBABILITY AND LOGIT UNCERTAINTY

Suppose $P = \mathrm{softmax}(L)$, define $\bar{L}_k = \mathbb{E}[L \mid \mathcal{G}_k]$, $U_k = L - \bar{L}_k$, and let $\Pi = I - C^{-1}\mathbf{1}\mathbf{1}^\top$.

Theorem 3 (Probability–logit variance sandwich). *If the quantities below are finite, then*

$$\frac{1}{2}\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(P \mid \mathcal{G}_k) \le \Delta_k \le \mathbb{E} D_{\mathrm{KL}}\big(P \big\| \mathrm{softmax}(\bar{L}_k)\big) \le \frac{1}{4}\mathbb{E}\|\Pi U_k\|_2^2. \tag{9}$$

If every softmax coordinate along the relevant conditional logit segments is at least $\rho > 0$, then

$$\frac{C^2 \rho^4}{2}\mathbb{E}\|\Pi U_k\|_2^2 \le \Delta_k. \tag{10}$$

This sandwich connects the information limit to variation across compatible worlds. The lower-bound constant is deliberately global and may be loose. We use it only as a diagnostic inequality, not as a tight practical certificate; the geometry-sensitive characterization below carries the mechanistic interpretation.

4.3 Fisher–Green characterization

Scale only unresolved logits by $L_s = \bar{L}_k + sU_k$, where $\mathbb{E}[U_k \mid \mathcal{G}_k] = 0$. Set $P_s = \mathrm{softmax}(L_s)$, $Q_k(s) = \mathbb{E}[P_s \mid \mathcal{G}_k]$, $d_k(s) = \mathbb{E}[D_{\mathrm{KL}}(P_s \| Q_k(s)) \mid \mathcal{G}_k]$, and $\Delta_k(s) = \mathbb{E}[d_k(s)]$. Let $p_k = \mathrm{softmax}(\bar{L}_k)$ and $J_k = \mathrm{diag}(p_k) - p_k p_k^\top$.

Theorem 4 (Fisher-weighted conditional covariance)**.** *Define the context-wise quantity* $\rho_k = \min_c p_{k,c} > 0$. *Assume* $\mathbb{E}\|U_k\|_2^3 < \infty$ *and*

$$\Xi_k := \mathbb{E}\left[\frac{\{\mathbb{E}[\|U_k\|_2^2 \mid \mathcal{G}_k]\}^2}{\rho_k}\right] < \infty.$$

For $0 \le s \le 1$,

$$d_k(s) = \frac{s^2}{2}\,\mathrm{tr}[J_k\,\mathrm{Cov}(U_k \mid \mathcal{G}_k)] + r_k(s), \tag{11}$$

For constants $a_C, b_C < \infty$ *depending only on* C, *the conditional remainder has an explicit moment bound in Eq. (45) of Appendix B. Consequently,*

$$\Delta_k(s) = \frac{s^2}{2}\mathbb{E}\,\mathrm{tr}[J_k\,\mathrm{Cov}(U_k \mid \mathcal{G}_k)] + \mathcal{R}_k(s), \tag{12}$$

with

$$|\mathcal{R}_k(s)| \le a_C s^3 \mathbb{E}\|U_k\|_2^3 + b_C s^4 \Xi_k. \tag{13}$$

Moreover,

$$\mathbb{E}D_{\mathrm{KL}}(Q_k(s)\|p_k) = O(s^4), \tag{14}$$

where the hidden constant is bounded by $d_C^2 \Xi_k$ *for a dimension-dependent* d_C. *A uniform* $p_{k,c} \ge \rho > 0$ *together with a finite fourth moment is a simpler, stronger sufficient condition, but is not required. Thus replacing the exact projection by* $\mathrm{softmax}(\bar{L}_k)$ *changes no second-order term.*

Dividing the remainder by the leading term gives a law-dependent finite-scale relative-error bound (Corollary 1; Appendix B.6). A binary-chain walkthrough in Appendix B.7 makes the influence–uncertainty–curvature product concrete.

The leading term is curvature-weighted unresolved uncertainty. For a class-centered covariance Γ with $\mathrm{tr}\,\Gamma = t$, it lies between $t\lambda^+_{\min}(J_k)/2$ and $t\lambda_{\max}(J_k)/2$. The upper endpoint is attained by rank-one uncertainty along the top Fisher eigenvector. Raw unresolved logit variance is therefore insufficient: its direction relative to predictive curvature matters.

4.4 Computable locality certificates

Let M be nonnegative with $\|M\|_2 \le 1$, H be base node logits, and $L_\infty = \sum_{r\ge 0} a_r M^r H$ for an absolutely summable sequence. Writing $L_k = \sum_{r=0}^k a_r M^r H$ and $\tau_k = \sum_{r>k} |a_r|$, a uniform root over n nodes satisfies

$$\overline{\Delta}_k \le \frac{1}{4n}\mathbb{E}\|(L_\infty - L_k)\Pi\|_F^2 \le \frac{\tau_k^2}{4n}\mathbb{E}\|H\Pi\|_F^2. \tag{15}$$

For PPR, set $\beta = 1 - \alpha$ and

$$G_\alpha = \alpha(I - \beta M)^{-1}, \qquad L_k = \alpha \sum_{r=0}^k \beta^r M^r H. \tag{16}$$

The uniform envelope decays as $\beta^{2(k+1)}$. If M is symmetric with eigenpairs (λ_j, u_j), the sharper certificate is

$$\frac{\alpha^2}{4n}\sum_{j=1}^n \frac{(\beta\lambda_j)^{2(k+1)}}{(1-\beta\lambda_j)^2}\|u_j^\top H\Pi\|_2^2. \tag{17}$$

Identification becomes graph-specific after a completion law is declared. Fix a root v, observe H_u for $u \in B_k(v)$, and let $\Sigma_{uu'|k} = \mathrm{Cov}(H_u, H_{u'} \mid \mathcal{G}_k)$. Then

$$\Gamma_{v,k} := \mathrm{Cov}(L_{\infty,v} \mid \mathcal{G}_k) = \sum_{u,u' \notin B_k(v)} (G_\alpha)_{vu}(G_\alpha)_{vu'}\Sigma_{uu'|k}. \tag{18}$$

Combining Eqs. (11) and (18) gives the context-wise relation

$$d_{v,k}(s) = \frac{s^2}{2}\,\mathrm{tr}(J_{v,k}\Gamma_{v,k}) + r_{v,k}(s). \tag{19}$$

Up to the controlled remainder, the second-order term is governed jointly by graph influence, remote dependence, and predictive sensitivity.

The envelope is rate-sharp in a worst-case construction. Choose two locally identical completions with equal probability, common root log-odds zero, and remote class-centered binary logits $\pm(a/2, -a/2)$ at node u. Here a is the magnitude of the remote log-odds, so the propagated root log-odds are $\pm a(G_\alpha)_{vu}$. This construction satisfies

$$\Delta_{v,k} \geq \frac{1}{2}\tanh^2\left(\frac{|a(G_\alpha)_{vu}|}{2}\right). \tag{20}$$

On a directed chain whose ambiguous node is at distance $k+1$, the lower and upper bounds are both $\Theta(\beta^{2k})$.

4.5 Nonlinear teachers and finite learners

Let a one-hop update satisfy $\|\Phi_H(X) - \Phi_H(Y)\|_F \leq q\|X - Y\|_F$, $q < 1$, and let X_∞ be its fixed point. If $X_{k+1,V}$ is radius-k measurable, then

$$\overline{\Delta}_k \leq \frac{1}{4n}\mathbb{E}\|(X_\infty - X_{k+1})\Pi\|_F^2 \leq \frac{q^{2(k+1)}}{4n(1-q)^2}\mathbb{E}\|X_1 - X_0\|_F^2. \tag{21}$$

Conditional accounting and this certificate extend beyond linear diffusion, whereas the explicit Green characterization does not. For any learned $\widehat{S}_k$, Eq. (3) remains exact. To estimate Δ_k, use M compatible completions to form $\widehat{Q}_M$ and N independent completions for evaluation:

$$\mathbb{E}\widehat{\Delta}_{M,N} = \Delta_k + \mathbb{E}D_{\mathrm{KL}}(Q_k\|\widehat{Q}_M). \tag{22}$$

The estimator is upward biased and consistent. Cross-fitting prevents resubstitution from disguising finite-M error.

5 Experiments

5.1 Questions and protocol

Following identification, characterization, and realization, we ask whether Δ_k is estimable under a declared law (RQ1), where useful shells enter (RQ2), what sets the boundary (RQ3), how much of the residual is learner-dependent (RQ4), and how the law changes it (RQ5).

The primary RQ1/RQ4 audit evaluates all KL components at matched observed contexts on fixed real graphs. On Cora, PubMed, and Roman-empire, we train a ten-step APPNP teacher with a width-64 encoder. The official validation nodes are split into teacher/student early stopping, law selection, and untouched predictive audit; test nodes are excluded. A matrix-normal law with row-standardized factor $I + \gamma M$ is fitted to training-node pre-propagation logits. Selection calibrates covariance and retains candidates within 0.01 conditional NLL per dimension of the best; this descriptive tolerance does not assume independent nodes. The full γ grid and untouched NLL/coverage audit are in Appendix C.1. Width-64 and width-256 MLP students receive the same propagated summaries at $k = 0, \ldots, 4$. Three trained teachers, 32 held-out roots each, and independent 16,384-draw scrambled Sobol batches for Q_k and KL evaluation support the decomposition. Let $\mathcal{U} = \mathcal{T} \cup \mathcal{V}_{\mathrm{es}} \cup \mathcal{V}_{\mathrm{law}}$. The fixed post-training design D_{fit} contains trained parameters, fitted law, and selected hyperparameters; $A_{\mathrm{fit}} = \sigma(D_{\mathrm{fit}})$ has no random coordinates after conditioning. Define

$$A_{\mathrm{obs}} = \sigma(A_{\mathrm{fit}}, H_{\mathcal{U}})\,. \tag{23}$$

We treat A_{fit} and A_{obs} as defining distinct explicit estimands. The primary projection conditions on (A_{obs}, Z_k); fit-only is a sensitivity analysis. Screening selects a working completion model for conditional accounting, not a unique data-generating law; held-out NLL, coverage, and target-aligned checks assess predictive adequacy. This post-encoder analysis fixes topology, raw features, and trained artifacts while completing H under the declared law. Students use local summaries of H. Student width changes a fixed artifact, not the random conditioning field. At each observed $g = (A_{\text{obs}}, Z_k)$, $t_k(g) = d_k(g) + e_k(g)$; Table 4 averages these conditional terms over fixed roots, not law-resampled Z_k. The field A_{obs} is not a full optimization transcript; A_{full} is audited separately. Further controls and numerical budgets are in Appendix C.1.

For the broader RQ1–RQ3 mechanism audit, we fix Cora, PubMed, Roman-empire, and Amazon-ratings with public splits (Yang et al., 2016; Pei et al., 2020; Platonov et al., 2023). A supervised feature MLP produces centered base logits. Training-node logits are used to calibrate a mean μ and covariance Σ. The primary completion law holds the real topology fixed and draws $H_u \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. For each of 32 test roots, one anchor is fixed inside $B_k(v)$ and remote logits are resampled. We evaluate $\alpha \in \{0.1, 0.2\}$ and $k \in \{0, 1, 2, 4, 8\}$. Each of 1,280 contexts uses 16,384 completions to estimate $\widehat{Q}_k$ and 16,384 independent completions for KL evaluation.

The controlled RQ4 audit additionally trains width-64 MLP students on the formal conditional population, using only Z_k, three sample sizes, and five seeds. Gauss–Hermite quadrature evaluates Q_k; 8,192 held-out contexts and 64 independent completions per context evaluate all three terms. A matched audit then spans four datasets, five splits, three seeds, four teleport probabilities, and six radii, comparing hop-MLPs, local GNNs, copied-parameter controls, and feature-only distillation controls. RQ5 changes only the completion law. Alongside i.i.d. nodes, it uses the standardized graph-correlated kernel

$$K_\gamma = D_\gamma^{-1/2}(I + \gamma M)^2 D_\gamma^{-1/2}, \qquad \gamma \in \{0.2, 0.5\}, \tag{24}$$

where $D_\gamma = \operatorname{diag}((I + \gamma M)^2)$. All three laws have the same nodewise marginal Σ, roots, PPR teachers, radii, and Monte Carlo budget. Crucially, the same i.i.d. anchor is reused under every law, so the observed logits in each local set are identical. This estimates law-specific $d_k^{(\gamma)}$ at matched contexts. Their average $\bar{d}_k^{(\gamma|\text{i.i.d.})} = \mathbb{E}_{Z_k \sim \text{i.i.d.}}[d_k^{(\gamma)}(Z_k)]$ is a standardized sensitivity estimand, not the correlated law's native Δ_k. Gaussian conditioning is computed without materializing K_γ. Appendix C gives complete conditioning formulae and a companion law-native population analysis.

6 Results

6.1 RQ1: The information limit can be estimated under a declared law

Under the primary i.i.d. law, the cross-fitted radius-zero gaps, averaged over teleport settings, are 0.00615, 0.0540, 0.0292, and 0.122 nats on Amazon-ratings, Cora, PubMed, and Roman-empire. At radius four, they fall to 1.76×10^{-6}, 4.55×10^{-5}, 1.17×10^{-5}, and 3.12×10^{-4}. None of 1,024 adjacent radius comparisons increases by more than 5×10^{-4}. This is an empirical pathwise audit, not a consequence of population monotonicity.

Because its completion law is explicit, each conditional quantity is identified under that law; the reported rootwise average describes the sampled contexts rather than a law-free graph property. In a separate controlled shell population, we compute Δ_k exactly and verify the cross-fit bias predicted by Eq. (22); it appears in Appendix D.

6.2 RQ2: Useful information enters unevenly across shells

The mean information gains over intervals $0 \to 1$, $1 \to 2$, $2 \to 4$, and $4 \to 8$ are 0.0402, 0.0101, 0.00247, and 9.3×10^{-5} nats, respectively. No interval gain falls below -5×10^{-4}. The decline is highly nonuniform, so a hop count alone does not express how much teacher-relevant information a shell contributes.

At $\varepsilon = 10^{-4}$, mean selected radii are 2.70, 4.75, 3.59, and 6.31 for Amazon-ratings, Cora, PubMed, and Roman-empire. All 256 root–teleport contexts reach the target by radius eight. The context

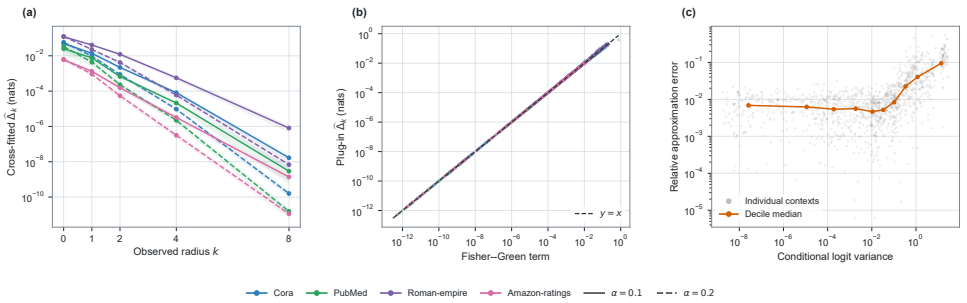


Figure 2: **Identifying and characterizing the information boundary.** (a) Cross-fitted $\widehat{\Delta}_k$ under the declared i.i.d. completion law; bands are descriptive 95% root-cluster intervals within fixed graphs, not graph-population confidence intervals. (b) The entropy plug-in estimate $\widehat{d}_{v,k}$ versus $\frac{1}{2}\,\mathrm{tr}(J_{v,k}\Gamma_{v,k})$ over 1,280 contexts; the dashed identity line is not fitted. (c) Relative second-order error versus exact conditional logit variance; orange points are decile medians.

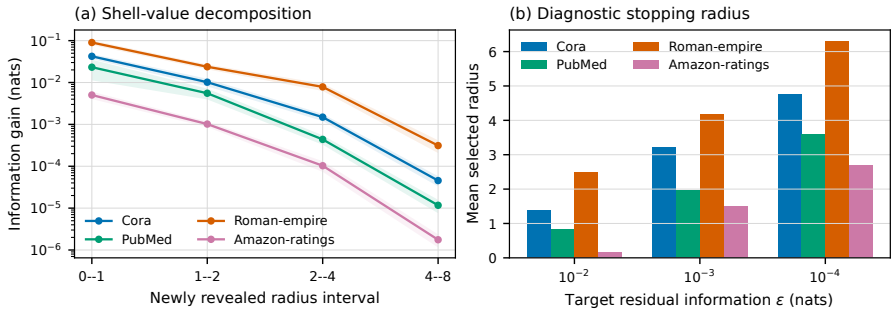


Figure 3: **Information-valued receptive fields.** (a) Differences of aggregated plug-in residuals, estimating interval shell values from Eq. (6), averaged over roots and teleport settings; bands are descriptive 95% root-level intervals within fixed graphs, not graph-population confidence intervals. (b) Mean first measured radius with conditional residual $\widehat{d}_k \le \varepsilon$, namely $\widehat{\tau}_{\varepsilon,8}$. All contexts meet all three targets by $k = 8$.

stopping radius is therefore node- and population-dependent, not a universal architectural depth. It is distinct from the single population design radius $k^*_{\mathrm{pop}}(\varepsilon)$.

6.3 RQ3: Influence, Uncertainty, and Curvature Jointly Set the Boundary

The Fisher–Green term has a Spearman correlation of 0.9999 with the plug-in gap over all 1,280 primary contexts. Its median relative error is 0.84% overall and 0.63% in the smallest-variance quartile, where the exact conditional variance is at most 1.89×10^{-4}. The approximation error decreases with conditional variance, consistent with the context-wise remainder characterization in Eq. (13). The coarser exact Green variance has a correlation of 0.978 with the cross-fitted gap. The probability lower bound, candidate KL, and logit upper bound show no violations above the 5×10^{-4} tolerance. The matched-context correlated-law audit retains correlations of 0.9999 and median relative errors below 0.86% under both graph-correlated laws (Table 10), supporting Fisher–Green as a mechanism-level explanation rather than a generic variance proxy.

Mechanism-specific audits support the certificate layer. On the directed chain, fitted slopes recover $2\log(1-\alpha)$ with maximum absolute error 1.38×10^{-4}. All 240 global-PPR teachers converge, and every pointwise tail inequality holds over 2,880 method rows. All 240 nonlinear fixed-point iterations also converge, with no analytic iterate or contraction-envelope violation. These audits verify the locality certificates and their implementations.

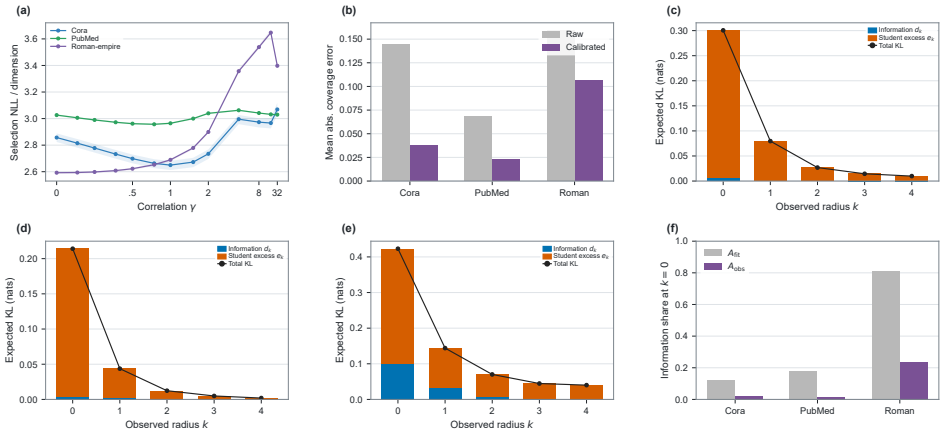


Figure 4: **Observed-context fitted-law decomposition.** (a) Selection NLL; (b) untouched-node coverage error; (c–e) Cora, PubMed, and Roman-empire conditional components averaged over fixed (A_{obs}, Z_k) test contexts, not resampled observations; (f) information-share change without training-side anchors. Shading spans three seeds; maximum independent QMC closure error is 2.53×10^{-4} nats. These panels report conditional quantities under the screened working law, not a law-free graph-population functional.

6.4 RQ4: INFORMATION AVAILABILITY DOES NOT GUARANTEE FINITE REALIZATION

Under the primary screened A_{obs} law, fitted-law decomposition gives radius-zero information shares of 2.3%, 1.7%, and 23.7% on Cora, PubMed, and Roman-empire (Figure 4). Oracle-relative excess (the realization gap) dominates all residuals; Roman-empire retains more missing information. By radius four, information accounts for less than 0.5% of total conditional KL (Table 4).

Fit-only radius-zero observed-context shares are 12.5%, 17.9%, and 80.9%; conditioning on training-side logits reduces them to 2.3%, 1.7%, and 23.7%, reversing Roman-empire's diagnosis.

The expanded γ search has no boundary optimum; calibration improves fit, although Roman-empire retains upper-tail undercoverage. Raw covariance gives A_{obs} shares of 3.8%, 2.0%, and 34.9%. Under the same law and inputs, width 256 reduces student excess by 7–51% at radii 2–4, with some smaller-radius regressions. Matched controls isolate realization and changes to A (Appendix C.1, Table 17, Figure 7, Table 19).

Under the same A_{obs} law and roots, the contractive map $X = H + q\tanh(MX)$, $q \in \{0.4, 0.7\}$, gives direct three-term accounting. Across Cora and Roman-empire, the maximum closure error is 1.40×10^{-3} nats (95th percentile 6.97×10^{-5}); truncation is below 10^{-6} (Appendix C.6, Figure 6).

6.5 RQ5: SENSITIVITY TO THE COMPLETION LAW

Because Δ_k is law-dependent, RQ5 separates quantitative sensitivity from qualitative stability. Matched-context kernel changes raise mean d_0 by 17–64%, while a Student-t_5 alternative changes it by at most 8% (Figure 9). Screening retains only Gaussian on Cora and both families elsewhere. Across 36 Roman-empire seed–law pairs, shares are 61.9–83.9% under A_{fit} versus 20.5–30.8% under A_{obs}; Cora and PubMed remain realization-dominated (Table 11).

Target-aligned 95% coverage at $k = 4$ is only 40.6%, 9.4%, and 79.2%, limiting law-free interpretation but not conditional accounting. Across retained laws, conditioning on training-side information shifts Roman-empire from information-dominated under A_{fit} to realization-dominated under A_{obs}; Cora and PubMed remain realization-dominated (Appendix C.5, Table 3).

Discussion. The framework separates information limits from finite-student excess and values shells exactly; Fisher–Green links influence, uncertainty, and curvature. Predictive checks assess fitted-law adequacy, while cross-law stability supports qualitative diagnoses; the nonlinear audit does not establish a general nonlinear Green representation.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, including language editing, manuscript organization, and LaTeX formatting; to retrieve and discover related work; and to draft sections of the manuscript. The authors reviewed and revised all AI-assisted text, independently verified the relevance and bibliographic details of retrieved references, and take full responsibility for the final manuscript, including its claims, citations, and presentation.

ETHICS STATEMENT

This computational study uses public benchmark data and involves no human participants or private data. A completion law should not be mistaken for the deployment population; every reported population quantity therefore names the law under which it is identified.

REPRODUCIBILITY STATEMENT

Appendix B gives complete proofs, and Appendix C specifies data splits, conditional populations, sampling, optimization, and uncertainty units. The supplementary artifact includes executable scripts, raw and aggregated results, run metadata, and deterministic seeds for all reported experiments.

REFERENCES

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=i80OPhOCVH2.

Justin Baker, Qingsong Wang, Cory D. Hauck, and Bao Wang. Implicit graph neural networks: A monotone operator viewpoint. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1521–1548, 2023. URL https://proceedings.mlr.press/v202/baker23a.html.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005. URL https://www.jmlr.org/papers/v6/banerjee05b.html.

Pietro Caputo, Zongchen Chen, Yuzhou Gu, and Yury Polyanskiy. Entropy contractions in markov chains: Half-step, full-step and continuous-time. *arXiv preprint arXiv:2409.07689*, 2024. URL https://arxiv.org/abs/2409.07689.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. doi: 10.1002/047174882X.

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Liò, and Michael Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7865–7887, 2023. URL https://proceedings.mlr.press/v202/di-giovanni23a.html.

Amir Eskandari, Aman Anand, Elyas Rashno, and Farhana Zulkernine. InfGraND: An influence-guided GNN-to-MLP knowledge distillation. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=lfzHR3YwlD.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://dl.acm.org/doi/10.5555/3294771.3294869.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

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

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

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

Frank Nielsen. Approximation and bounding techniques for the Fisher–Rao distances between parametric statistical models. In *Handbook of Statistics*, volume 51, pp. 67–116. Elsevier, 2024. doi: 10.1016/bs.host.2024.06.003.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1e2agrFvS.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tJbbQfw-5wv.

Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, volume 161 of *The IMA Volumes in Mathematics and its Applications*, pp. 211–249. Springer, 2017. doi: 10.1007/978-1-4939-7005-6_8.

Jan Seyfried, Sayantan Sen, and Marco Tomamichel. Testing (conditional) mutual information. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5246–5247, 2025. URL https://proceedings.mlr.press/v291/seyfried25a.html.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V. Chawla. Learning MLPs on graphs: A unified view of effectiveness, robustness, and efficiency. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Cs3r5KLdoj.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042.

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

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

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

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

Ling Yang, Ye Tian, Minkai Xu, Zhongyi Liu, Shenda Hong, Wei Qu, Wentao Zhang, Bin Cui, Muhan Zhang, and Jure Leskovec. VQGraph: Rethinking graph representation space for bridging GNNs and MLPs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48, 2016. URL https://proceedings.mlr.press/v48/yanga16.html.

Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and En-Hui Yang. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=yV6wwEbtkR.

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

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old MLPs new tricks via distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=4p6_5HBWPCw.

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

A Notation and Estimands

Table 1: Quantities that must remain distinct in theory and experiment.

Symbol	Meaning
d_k	Conditional residual $\mathbb{E}[D_{\mathrm{KL}}(P\|Q_k) \mid \mathcal{G}_k]$, measurable from the current context under a declared completion law.
Δ_k	Population information limit $\mathbb{E}[d_k]$.
$t_k(g), e_k(g)$	Conditional total KL and student excess at a realized context g; $t_k(g) = d_k(g) + e_k(g)$.
$\bar{t}_k^{\mathrm{obs}}, \bar{d}_k^{\mathrm{obs}}, \bar{e}_k^{\mathrm{obs}}$	Averages of conditional components over fixed held-out test-root contexts; not in general population expectations over resampled local observations.
$k^*_{\mathrm{pop}}(\varepsilon)$	Smallest fixed population radius whose mean residual is at most ε.
$\tau_{\varepsilon,K}$	First context-dependent threshold crossing on a finite radius grid, truncated at K; threshold failure is recorded.
$\mathcal{E}_k$	Learner-dependent realization gap $\mathbb{E}D_{\mathrm{KL}}(Q_k\|S_k)$.
$r_k(s), \mathcal{R}_k(s)$	Conditional and population Taylor remainders in the Fisher expansion, not realization errors or diffusion tails.
ρ_k	Context-wise minimum coordinate of $p_k = \mathrm{softmax}(\bar{L}_k)$; it is positive for every finite context but need not have a uniform population lower bound.
T_k	Deterministic truncation tail $L_\infty - L_k$.

B Complete Proofs

B.1 Proof of Theorem 1

For every class c, add and subtract $\log Q_{k,c}$:

$$\mathbb{E}D_{\mathrm{KL}}(P\|S_k) = \mathbb{E}\sum_c P_c \log \frac{P_c}{Q_{k,c}} + \mathbb{E}\sum_c P_c \log \frac{Q_{k,c}}{S_{k,c}}. \tag{25}$$

The logarithmic ratio in the second term is $\mathcal{G}_k$-measurable. The tower property therefore gives

$$\mathbb{E}\left[P_c \log \frac{Q_{k,c}}{S_{k,c}}\right] = \mathbb{E}\left[\mathbb{E}[P_c \mid \mathcal{G}_k] \log \frac{Q_{k,c}}{S_{k,c}}\right] \tag{26}$$

$$= \mathbb{E}\left[Q_{k,c} \log \frac{Q_{k,c}}{S_{k,c}}\right]. \tag{27}$$

Summing over classes in Eq. (25) proves the identity. Both terms are nonnegative, so Q_k is almost surely optimal and is unique wherever the conditional distribution has support.

B.2 Proof of the information and increment identities

By construction, the conditional law of $\widetilde{Y}$ given (W, V, A) is P. Its law given (A, Z_k) is $\mathbb{E}[P \mid A, Z_k] = Q_k$. Hence

$$I(\widetilde{Y}; W, V \mid A, Z_k) = H(\widetilde{Y} \mid A, Z_k) - H(\widetilde{Y} \mid W, V, A, Z_k) \tag{28}$$

$$= \mathbb{E}H(Q_k) - \mathbb{E}H(P) \tag{29}$$

$$= \mathbb{E}\sum_c P_c \log \frac{P_c}{Q_{k,c}} = \Delta_k. \tag{30}$$

Because observations are nested, $Q_k = \mathbb{E}[Q_{k+1} \mid \mathcal{G}_k]$. Applying Theorem 1 with P, Q_{k+1}, and Q_k yields

$$\Delta_k = \Delta_{k+1} + \mathbb{E}D_{\mathrm{KL}}(Q_{k+1}\|Q_k). \tag{31}$$

The same projection identity localized to $\mathcal{G}_k$ gives

$$d_k = \mathbb{E}[D_{\mathrm{KL}}(P\|Q_{k+1}) \mid \mathcal{G}_k] + \mathbb{E}[D_{\mathrm{KL}}(Q_{k+1}\|Q_k) \mid \mathcal{G}_k] \tag{32}$$

$$= \mathbb{E}[d_{k+1} + D_{\mathrm{KL}}(Q_{k+1}\|Q_k) \mid \mathcal{G}_k], \tag{33}$$

which proves Eq. (5). Taking expectations recovers the population increment identity. Applying the first part of this proof to the two nested information sets gives $\mathbb{E}D_{\mathrm{KL}}(Q_{k+1}\|Q_k) = I(\widetilde{Y}; Z_{k+1} \mid A, Z_k)$.

B.3 Proof of Theorem 2

Summing the one-step identity $\Delta_r - \Delta_{r+1} = \mathbb{E}D_{\mathrm{KL}}(Q_{r+1}\|Q_r)$ from $r = k$ to $j - 1$ proves the first equality in Eq. (6); the conditional-information identity above gives the second. If $\bigvee_r \mathcal{G}_r$ reveals P, martingale convergence gives $Q_r \to P$ and $\Delta_r \downarrow 0$, so monotone convergence yields the infinite shell sum.

For the stopping statement, the pointwise telescoping identity is

$$D_{\mathrm{KL}}(P\|Q_0) - D_{\mathrm{KL}}(P\|Q_\tau) = \sum_{r=0}^{K-1} \mathbf{1}\{\tau > r\} \left[D_{\mathrm{KL}}(P\|Q_r) - D_{\mathrm{KL}}(P\|Q_{r+1})\right]. \tag{34}$$

Because τ is adapted, $\mathbf{1}\{\tau > r\}$ is $\mathcal{G}_r$-measurable. The localized projection identity therefore gives

$$\mathbb{E}[\mathbf{1}\{\tau > r\} \{D_{\mathrm{KL}}(P\|Q_r) - D_{\mathrm{KL}}(P\|Q_{r+1})\}] \tag{35}$$

$$= \mathbb{E}[\mathbf{1}\{\tau > r\} D_{\mathrm{KL}}(Q_{r+1}\|Q_r)]. \tag{36}$$

Taking expectations and summing proves Eq. (7). Finally, $Q_\tau = \mathbb{E}[P \mid \mathcal{G}_\tau]$ follows by checking the defining conditional-expectation property on each event $\{\tau = r\}$.

B.4 Proof of Proposition 1

Let $p_i = T(w_i, v)$, and let the two worlds share $(A, Z_k) = (a, z)$. Under $\mu_0 = \delta_{w_0}$, the conditional teacher output is deterministic, so $Q_k = p_0$ and $\Delta_{v,k} = 0$. For any $\varepsilon \in (0, 1)$, define $\mu_1 = (1-\varepsilon)\delta_{w_0} + \varepsilon\delta_{w_1}$. The conditional mean is $q = (1-\varepsilon)p_0 + \varepsilon p_1$, and

$$\Delta_{v,k} = (1-\varepsilon)D_{\mathrm{KL}}(p_0\|q) + \varepsilon D_{\mathrm{KL}}(p_1\|q) > 0$$

because $p_0 \neq p_1$. Both laws put probability one on the same observable $(A, Z_k) = (a, z)$, so its distribution cannot identify the gap. The complete-world distributions do differ. Nevertheless, w_0 has positive probability under both; a statistic of one complete realization cannot recover both target values without error uniformly over these laws.

B.5 Proof of Theorem 3

Let $Q = \mathbb{E}[P \mid \mathcal{G}_k]$. Pinsker and $\|x\|_1 \geq \|x\|_2$ give

$$\Delta_k \geq \frac{1}{2}\mathbb{E}\|P - Q\|_2^2 = \frac{1}{2}\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(P \mid \mathcal{G}_k).$$

The logit conditional mean $\bar{L}_k$ is $\mathcal{G}_k$-measurable, so the projection property yields

$$\Delta_k \leq \mathbb{E}D_{\mathrm{KL}}\big(P\big\|\mathrm{softmax}(\bar{L}_k)\big)\,.$$

Let $\psi(a) = \log\sum_c e^{a_c}$. The categorical exponential-family identity gives

$$D_{\mathrm{KL}}(\mathrm{softmax}(a+b)\|\,\mathrm{softmax}(a)) = D_\psi(a, a+b). \tag{37}$$

For $p = \mathrm{softmax}(a)$, $\nabla^2\psi(a) = \mathrm{diag}(p) - pp^\top$. For any unit u, its quadratic form equals the variance of a variable that takes value u_c with probability p_c. Popoviciu's inequality gives

$$u^\top\nabla^2\psi(a)u \leq \frac{(\max_c u_c - \min_c u_c)^2}{4} \leq \frac{1}{2}\|u\|_2^2. \tag{38}$$

Thus ψ is $1/2$-smooth and its Bregman divergence is at most $\|b\|_2^2/4$. Softmax is invariant to adding a multiple of $\mathbf{1}$, so, with $a = \bar{L}_k$ and $b = U_k$, the perturbation may be replaced by ΠU_k. Taking expectations proves the upper half of Eq. (9).

For the bounded-support lower bound, take conditionally i.i.d. logits L, L' and write $P = \mathrm{softmax}(L)$, $P' = \mathrm{softmax}(L')$. For every class-centered x, the softmax Jacobian $J(p)$ satisfies

$$x^\top J(p)x = \frac{1}{2}\sum_{i,j} p_i p_j (x_i - x_j)^2 \geq C\rho^2\|x\|_2^2.$$

Integrating the Jacobian along the segment between L and L' and using Cauchy–Schwarz gives $\|P - P'\|_2 \geq C\rho^2\|\Pi(L - L')\|_2$. Conditional pairwise-variance identities then yield

$$\mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(P \mid \mathcal{G}_k) \geq C^2\rho^4\mathbb{E}\|\Pi(L - \bar{L}_k)\|_2^2.$$

Combining with the probability-variance lower bound proves Eq. (10).

B.6 Proof of Theorem 4

Condition on $\mathcal{G}_k$, abbreviate $m = \bar{L}_k$, $U = U_k$, and let $\psi(a) = \log\sum_c e^{a_c}$. The exponential-family identity and a third-order Taylor expansion give

$$\mathbb{E}D_{\mathrm{KL}}(\mathrm{softmax}(m + sU)\|\,\mathrm{softmax}(m)) = \frac{s^2}{2}\mathbb{E}[U^\top J_k U \mid \mathcal{G}_k] + R_{\mathrm{cand}}(s) \tag{39}$$

$$= \frac{s^2}{2}\,\mathrm{tr}[J_k\,\mathrm{Cov}(U \mid \mathcal{G}_k)] + R_{\mathrm{cand}}(s). \tag{40}$$

For a direction u, the third derivative of log-sum-exp is the third centered moment of the categorical variable u_X. Since $|u_X - \mathbb{E}u_X| \leq 2\|u\|_2$, its absolute value is at most $8\|u\|_2^3$. Integrating the directional Hessian along the path gives the conservative explicit choice $a_C = 8/3$, such that

$$|R_{\mathrm{cand}}(s)| \leq a_C s^3\mathbb{E}[\|U\|_2^3 \mid \mathcal{G}_k]. \tag{41}$$

Because the linear term vanishes under $\mathbb{E}[U \mid \mathcal{G}_k] = 0$, the softmax Hessian bound $\|D^2 \operatorname{softmax}(x)[u,u]\|_2 \le 8\|u\|_2^2$ permits the conservative choice $d_C = 4$, such that

$$\|Q_k(s) - p_k\|_2 \le d_C s^2 \mathbb{E}[\|U\|_2^2 \mid \mathcal{G}_k]. \tag{42}$$

For categorical distributions, $D_{\mathrm{KL}}(q\|p) \le \chi^2(q,p)$. At the fixed context, let $\rho_k = \min_c p_{k,c} > 0$. Then

$$D_{\mathrm{KL}}(Q_k(s)\|p_k) \le \rho_k^{-1}\|Q_k(s) - p_k\|_2^2 \tag{43}$$

$$\le \rho_k^{-1} d_C^2 s^4 \left(\mathbb{E}[\|U\|_2^2 \mid \mathcal{G}_k]\right)^2. \tag{44}$$

Taking expectations gives $\mathbb{E}D_{\mathrm{KL}}(Q_k(s)\|p_k) \le d_C^2 s^4 \Xi_k$, proving Eq. (14), including the stated context-integrated constant. The conditional projection identity subtracts this nonnegative correction from Eq. (40). Therefore

$$r_k(s) = R_{\mathrm{cand}}(s) - D_{\mathrm{KL}}(Q_k(s)\|p_k),$$

and Eqs. (41) and (44), with $b_C = d_C^2$, prove Eq. (11) and the almost-sure bound

$$|r_k(s)| \le a_C s^3 \mathbb{E}[\|U_k\|_2^3 \mid \mathcal{G}_k] + \frac{b_C}{\rho_k} s^4 \left(\mathbb{E}[\|U_k\|_2^2 \mid \mathcal{G}_k]\right)^2. \tag{45}$$

Finally define $\mathcal{R}_k(s) = \mathbb{E}[r_k(s)]$. Taking expectations under the assumed $\Xi_k < \infty$ proves Eqs. (12) and (13). A uniform lower bound $\rho_k \ge \rho > 0$ plus $\mathbb{E}\|U_k\|_2^4 < \infty$ implies this condition by conditional Jensen, but the Gaussian completion experiments use the context-wise statement: unbounded logits do not admit a population-uniform ρ.

For the extremal statement, diagonalize J_k on $\mathbf{1}^\perp$. If $\Gamma \succeq 0$ is supported on this subspace and $\operatorname{tr}\Gamma = t$, the Rayleigh–Ritz theorem yields

$$t\lambda_{\min}^+(J_k) \le \operatorname{tr}(J_k\Gamma) \le t\lambda_{\max}(J_k). \tag{46}$$

Taking $\Gamma = tuu^\top$ for a top eigenvector u attains the upper endpoint.

Corollary 1 (Law-dependent finite-scale relative-error bound)**.** *Let* $F_k = \frac{1}{2}\mathbb{E}\operatorname{tr}[J_k \operatorname{Cov}(U_k \mid \mathcal{G}_k)] > 0$. *For* $0 < s \le 1$, *Theorem 4 gives*

$$\left|\frac{\Delta_k(s)}{s^2 F_k} - 1\right| \le s\frac{a_C \mathbb{E}\|U_k\|_2^3}{F_k} + s^2\frac{b_C \Xi_k}{F_k}.$$

Consequently, the quadratic Fisher–Green term has relative error at most η *whenever the law-dependent right-hand side is at most* η.

The statement follows by dividing Eq. (13) by $s^2 F_k$. It is informative only when predictive curvature and unresolved variation make F_k non-negligible; this explicitly exposes the regime where a small absolute remainder could still have large relative error.

B.7 A BINARY CHAIN WALKTHROUGH FOR FISHER–GREEN

Consider a root v, one unresolved remote node u, and a binary teacher. Let $g = (G_\alpha)_{vu}$ be the Green influence from u to v, and let the unresolved centered logit be $U = g\xi(1,-1)^\top$, where $\mathbb{E}[\xi \mid \mathcal{G}_k] = 0$ and $\operatorname{Var}(\xi \mid \mathcal{G}_k) = \sigma^2$. If the conditional mean probability is $p_k = (p, 1-p)$, then

$$\Gamma_{v,k} = g^2\sigma^2\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad \frac{s^2}{2}\operatorname{tr}(J_k\Gamma_{v,k}) = 2s^2 g^2 \sigma^2 p(1-p).$$

The four factors are explicit: g^2 is squared graph influence, σ^2 is unresolved uncertainty, $p(1-p)$ is predictive curvature, and s^2 is perturbation scale. At $p = 1/2$ the leading gap is $s^2 g^2 \sigma^2/2$; near a confident prediction it shrinks even when remote variance is unchanged. On a directed PPR chain with u at distance $k+1$, $g = \alpha(1-\alpha)^{k+1}$, recovering the $(1-\alpha)^{2(k+1)}$ rate in one line. The theorem and experiments extend this calculation to multiple classes, remote nodes, and correlated uncertainty.

B.8 Proof of the PPR certificates

Upper certificate. The truncated logits $L_{k,v}$ are measurable with respect to the declared radius-k observation. Theorem 3 therefore gives

$$\Delta_{v,k} \le \mathbb{E} D_{\mathrm{KL}}(P_v \| \operatorname{softmax}(L_{k,v})) \le \frac{1}{4}\mathbb{E}\|T_{k,v}\Pi\|_2^2, \qquad T_k := L_\infty - L_k.$$

Averaging over a uniform root yields the first inequality in Eq. (15). Since $\|M\|_2 \le 1$,

$$\|T_k\Pi\|_F \le \alpha \sum_{r=k+1}^{\infty} \beta^r \|M^r H\Pi\|_F \tag{47}$$

$$\le \alpha \sum_{r=k+1}^{\infty} \beta^r \|H\Pi\|_F = \beta^{k+1}\|H\Pi\|_F. \tag{48}$$

Squaring and taking expectations proves the second inequality.

Spectral formula. For $M = U\Lambda U^\top$, summing the geometric tail gives

$$T_k = U \operatorname{diag}\left(\frac{\alpha(\beta\lambda_j)^{k+1}}{1-\beta\lambda_j}\right) U^\top H.$$

Orthogonal invariance of the Frobenius norm gives Eq. (17).

Conditional Green covariance. Condition on the fixed graph, the local logits, and all side information. The random remote part of the root logit is

$$L_{\infty,v} - \mathbb{E}[L_{\infty,v} \mid Z_k] = \sum_{u \notin B_k(v)} (G_\alpha)_{vu} \left(H_u - \mathbb{E}[H_u \mid Z_k]\right).$$

Taking the conditional covariance before imposing independence gives

$$\operatorname{Cov}(L_{\infty,v} \mid \mathcal{G}_k) = \sum_{u,u' \notin B_k(v)} (G_\alpha)_{vu}(G_\alpha)_{vu'} \operatorname{Cov}(H_u, H_{u'} \mid \mathcal{G}_k),$$

which is Eq. (18). Conditional independence removes all cross-covariance terms, leaving the special case

$$\operatorname{Cov}(L_{\infty,v} \mid Z_k) = \sum_{u \notin B_k(v)} (G_\alpha)_{vu}^2 \Sigma_u.$$

Multiplying by Π on both sides and taking the trace gives the exact class-centered conditional logit variance used in Eq. (9). If every $\Sigma_u = \Sigma$, the scalar factor is $\sum_{u \notin B_k(v)} (G_\alpha)_{vu}^2$.

Remote-completion lower certificate. Write $g = (G_\alpha)_{vu}$ and $t = ag$. Give the two compatible completions equal probability, root background log-odds zero, and remote class-centered logits $\pm(a/2, -a/2)$ at u; hence the root log-odds are $\pm t$. The teacher distributions at v are

$$P^{(+)} = (\sigma(t), 1-\sigma(t)), \qquad P^{(-)} = (\sigma(-t), 1-\sigma(-t)).$$

Their conditional mean is $(1/2, 1/2)$, and their total-variation distance is $\tanh(|t|/2)$. The auxiliary two-completion bound below, with equal branch probabilities, proves Eq. (20). If $|t| \le 2$, then $\tanh(|t|/2) \ge |t|/4$, so

$$\Delta_{v,k} \ge \frac{a^2 g^2}{32}.$$

For a nonnegative propagation operator and a node at distance r, the PPR series also gives

$$g \ge \alpha\beta^r (M^r)_{vu}.$$

Matching chain. On a directed shift chain with the ambiguous node at distance $k+1$, $(M^{k+1})_{vu} = 1$, all other powers linking u to v vanish, and $g = \alpha\beta^{k+1}$. The preceding lower bound becomes

$$\Delta_{v,k} \geq \frac{a^2\alpha^2}{32}\beta^{2(k+1)}.$$

The binary specialization of Theorem 3 gives

$$\Delta_{v,k} \leq \frac{a^2\alpha^2}{8}\beta^{2(k+1)}.$$

Thus the rate is $\Theta(\beta^{2k})$.

B.9 Proof of the global-filter certificate

The truncated global-filter logits are measurable with respect to the declared radius-k observation. Theorem 3 therefore gives the first inequality. Absolute summability and $\|M\|_2 \leq 1$ give

$$\|(L_\infty - L_k)\Pi\|_F \leq \sum_{r=k+1}^{\infty} |a_r| \|M^r H\Pi\|_F \tag{49}$$

$$\leq \left(\sum_{r=k+1}^{\infty} |a_r|\right) \|H\Pi\|_F = \tau_k \|H\Pi\|_F. \tag{50}$$

Squaring, averaging over graph worlds, and dividing by n proves Eq. (15).

B.10 Proof of the nonlinear certificate

Banach's fixed-point theorem gives a unique X_∞. The theorem's locality-aligned initialization assumption makes $X_{k+1,V}$ measurable with respect to the declared radius-k observation. Hence Theorem 3 gives

$$\overline{\Delta}_k \leq \frac{1}{4n}\mathbb{E}\|(X_\infty - X_{k+1})\Pi\|_F^2.$$

Moreover, contractivity gives

$$\|X_\infty - X_{k+1}\|_F \leq \sum_{j=k+1}^{\infty} \|X_{j+1} - X_j\|_F \tag{51}$$

$$\leq \sum_{j=k+1}^{\infty} q^j \|X_1 - X_0\|_F \tag{52}$$

$$= \frac{q^{k+1}}{1-q}\|X_1 - X_0\|_F. \tag{53}$$

Class centering cannot increase the norm. Squaring and averaging proves Eq. (21). For the audited residual update $X_{t+1} = H + q\tanh(MX_t)$, elementwise $\tanh$ is 1-Lipschitz, so the composition with M has Lipschitz constant at most q. Here $X_1 = H$, and induction shows that $X_{k+1,V}$ depends only on H within radius k; hence the locality-aligned condition holds.

B.11 Auxiliary two-completion lower bound

Pinsker's inequality gives $D_{\mathrm{KL}}(P^{(i)}\|Q) \geq 2\,\mathrm{TV}(P^{(i)}, Q)^2$. Since $Q = \pi_0 P^{(0)} + \pi_1 P^{(1)}$,

$$\mathrm{TV}(P^{(0)}, Q) = \pi_1 d, \qquad \mathrm{TV}(P^{(1)}, Q) = \pi_0 d. \tag{54}$$

The weighted sum is at least $2(\pi_0\pi_1^2 + \pi_1\pi_0^2)d^2 = 2\pi_0\pi_1 d^2$. When top-1 classes differ, a deterministic function of the shared local observation outputs the same class for both completions and must therefore disagree on at least one branch with mass at least $\min(\pi_0, \pi_1)$.

B.12 Finite-sample interval for conditional completions

For a fixed context $g = (A, Z_k)$, write $d_k(g) = \mathbb{E}[D_{\mathrm{KL}}(P\|Q_k) \mid g]$. Construct $\widehat{Q}_M = M^{-1}\sum_{i=1}^{M} P_i$ and evaluate it on an independent set of conditionally i.i.d. completions $P'_1, \ldots, P'_N$. Conditional independence and the same cross-term calculation as Theorem 1 give

$$\mathbb{E}[\widehat{d}_{M,N}(g) \mid g] = d_k(g) + \mathbb{E}[D_{\mathrm{KL}}(Q_k\|\widehat{Q}_M) \mid g]. \tag{55}$$

If $\widehat{Q}_{M,c} \geq \eta$, each evaluation KL lies in $[0, \log(1/\eta)]$. Hoeffding's inequality bounds evaluation error by

$$a_N = \log(1/\eta)\sqrt{\frac{\log(4/\delta)}{2N}}. \tag{56}$$

Coordinate-wise Hoeffding plus a union bound gives $\|Q_k - \widehat{Q}_M\|_2^2 \leq C\log(4C/\delta)/(2M)$. Since $D_{\mathrm{KL}}(q\|\widehat{q}) \leq \|q - \widehat{q}\|_2^2/\eta$, define

$$b_M = \frac{C\log(4C/\delta)}{2M\eta}. \tag{57}$$

Combining the two events yields, with probability at least $1 - \delta$,

$$\max\{0, \widehat{d}_{M,N}(g) - a_N - b_M\} \leq d_k(g) \leq \widehat{d}_{M,N}(g) + a_N. \tag{58}$$

This concentration statement requires i.i.d. conditional draws. Our scrambled Sobol evaluations instead report between-scramble numerical variation; their within-scramble points do not satisfy this Hoeffding premise.

C Additional Experimental Details

C.1 Fitted completion law and observed-context decomposition

Population object and data separation. We use Cora, PubMed, and Roman-empire split 0. We use the official training nodes to fit the teacher, students, and completion-law parameters. The official validation nodes are first split in half: the first half selects teacher/student epochs. From the second half we use all 398 Cora nodes and fixed 512-node PubMed and Roman-empire subsamples, then divide them again into law-selection and untouched predictive-audit sets (199/199 and 256/256 nodes). No validation set contains a test root. Results use 32 test roots drawn without replacement for each of three model seeds. Roots within one graph are dependent; intervals and seed ranges are therefore diagnostic uncertainty, not graph-level sampling confidence.

The teacher has a two-layer width-64 feature encoder followed by ten APPNP steps with teleport probability 0.1. It is trained end-to-end on hard labels with Adam, learning rate 0.01, weight decay 5×10^{-4}, dropout 0.35, at most 300 epochs, and patience 40. Let $H \in \mathbb{R}^{n\times(C-1)}$ be its class-centered pre-propagation logits in an orthonormal contrast basis. The graph topology and propagation operator remain fixed; the random object is the compatible pre-propagation logit field, not a new graph. These worlds are not claimed to arise from alternative raw-feature worlds passed through the fixed encoder. Raw features, labels, splits, topology, and retained learned parameters are fixed design variables in this conditional population. They therefore do not contribute additional random coordinates to the post-training estimand.

Training-fitted candidate laws and held-out screening. For $\gamma \in \{0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2, 4, 8, 16, 32\}$, define the row-standardized sparse factor

$$B_\gamma = R_\gamma^{-1}(I + \gamma M), \qquad (R_\gamma)_{uu} = \|(I + \gamma M)_{u,:}\|_2,$$

and the temperature-scaled matrix-normal candidate

$$H = \mathbf{1}\mu^\top + B_\gamma E, \qquad E_{u,:} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, t_\gamma\Sigma).$$

The mean μ and 10%-shrunk class covariance Σ are estimated using training nodes only. Row standardization preserves the same marginal class covariance at every node before temperature calibration, while γ changes graph dependence. For each γ, the law-selection nodes fit the scalar

Table 2: **Untouched predictive-adequacy diagnostics.** Correlation and covariance temperature are frozen before these nodes are evaluated. Coverage is joint Mahalanobis coverage at nominal .50/.80/.95.

Data	Law	γ	scale	NLL/dim.	Cov.50	Cov.80	Cov.95
Cora	Raw	1,1,1	1.000	2.723	0.807	0.901	0.975
Cora	Calibrated	1,1,1	0.604	2.682	0.543	0.799	0.881
PubMed	Raw	0.75,0.75,0.75	1.000	2.933	0.586	0.895	0.974
PubMed	Calibrated	0.75,0.75,0.75	0.845	2.920	0.531	0.827	0.960
Roman-empire	Raw	0,0,0	1.000	2.658	0.826	0.928	0.974
Roman-empire	Calibrated	0,0,0	0.589	2.603	0.660	0.767	0.826

Table 3: **Target-aligned held-out Gaussian predictive audit.** The random target is the trained teacher's class-contrast aggregate $g_v^\top H \mid (A_{\mathrm{obs}}, Z_k)$, rather than one node's $H_i \mid H_{\mathcal{T}}$. At each radius, 32 audit roots per seed (96 per graph) were selected by fixed independent seeds after freezing the screened Gaussian law and temperature. RMSE is the root-mean-square Euclidean error of the class-contrast logit vector. NLL and coverage are descriptive within dependent nodes of one graph; the fixed audit subset is not an independent graph-level validation sample.

Dataset	k	Mean NLL	Root-logit RMSE	50% joint	80% joint	95% joint
Cora	0	1.409	0.940	28.1%	44.8%	61.5%
	2	-1.035	0.318	7.3%	17.7%	20.8%
	4	-15.834	0.043	6.3%	15.6%	40.6%
PubMed	0	7.619	0.860	16.7%	19.8%	39.6%
	2	32.310	0.350	1.0%	6.3%	11.5%
	4	48.460	0.056	2.1%	4.2%	9.4%
Roman-empire	0	7.572	3.690	52.1%	68.8%	85.4%
	2	-16.090	0.937	56.3%	74.0%	78.1%
	4	-51.570	0.121	47.9%	75.0%	79.2%

Gaussian NLL minimizer $t_\gamma = \{\sum_i m_i\}/\{(C-1)n_{\mathrm{sel}}\}$, clipped to $[0.1, 4]$, where m_i is the unscaled conditional Mahalanobis statistic. The same nodes screen γ by paired conditional NLL. A candidate is retained when its mean NLL per class-contrast dimension is within 0.01 of that of the best candidate on the selection nodes. This prespecified descriptive tolerance avoids an invalid node-level i.i.d. standard-error calculation; retained laws are termed *validation-competitive*; this designation does not imply inferential equivalence. Only after these choices are frozen do untouched audit nodes report conditional NLL, RMSE, and joint Mahalanobis coverage. These diagnostics quantify predictive adequacy within the retained Gaussian family.

Teacher-target predictive audit. After freezing the screened Gaussian candidate, temperature, and selection rule, we draw 32 audit roots per teacher seed by an independent fixed RNG. At $k \in \{0, 2, 4\}$ we predict the class-contrast aggregate $g_v^\top H$ conditional on the same (A_{obs}, Z_k) used in the primary KL accounting. We evaluate conditional Gaussian NLL, Euclidean mean error, and 50/80/95% joint Mahalanobis coverage against the held-out trained teacher aggregate. These 96 roots per graph are dependent and form a fixed subset of the audit pool; no candidate is reselected against them. Table 3 exposes substantial undercoverage, especially at larger radii. The single-node audit does not certify the teacher-target conditional law, and the observed-context KL shares retain their fitted-law interpretation.

Declared conditioning sets and conditional test-root decomposition. At radius k, the student receives $[H, MH, \ldots, M^k H]$, hence is measurable with respect to the fixed-topology propagation observation. Boundary degrees and normalization weights needed for these summaries are included in Z_k; this is not an induced-subgraph-only or raw-feature-only graphless observation. A two-layer width-64 MLP is distilled from the frozen teacher using training nodes and selected on the early-stopping half of validation; a matched width-256 MLP provides a same-information

capacity intervention. We distinguish two compatible-world estimands. Let D_{fit} denote the frozen teacher/student parameters, fitted law parameters, and selected hyperparameters. In the ambient model we write $A_{\mathrm{fit}} = \sigma(D_{\mathrm{fit}})$, but the compatible-world law is evaluated conditional on the observed D_{fit}, so these artifacts add no random coordinates. Different student widths have different frozen parameters but the same random conditioning field. The primary coordinate-conditioned sigma-field $A_{\mathrm{obs}} = \sigma(A_{\mathrm{fit}}, H_{\mathcal{U}})$, with $\mathcal{U} = \mathcal{T} \cup \mathcal{V}_{\mathrm{es}} \cup \mathcal{V}_{\mathrm{law}}$, additionally fixes the realized pre-propagation logits on official training, early-stopping, and law-selection nodes. Untouched audit nodes and test roots are excluded. This is a coordinate-conditioned post-training estimand. It does not condition on every intermediate tensor observed during optimization. Nor is every coordinate in A_{obs} a literal input to the finite student; the student-excess component is relative to the declared conditional oracle, not optimization failure alone.

For a test root v, APPNP linearity writes the teacher logits as $g_v^\top H$. Gaussian conditioning first uses the training-side anchors in A_{obs}, then the newly revealed nodes in $B_k(v)$. The noiseless Schur complement defines the exact conditional moments. The primary decomposition uses a 10^{-6} linear-system ridge and the same lower bound on scalar conditional variance; a noiseless solver audit removes that artificial observation noise and output floor. Repeating the computation without the training-side anchors gives the fit-only sensitivity estimand (A_{fit}, Z_k). For each realized $g_{v,k} = (A_{\mathrm{obs}}, Z_{v,k})$, define

$$
\begin{aligned}
t_{v,k}(g) &= \mathbb{E}[D_{\mathrm{KL}}(P_v \| S_{v,k}) \mid g], \\
d_{v,k}(g) &= \mathbb{E}[D_{\mathrm{KL}}(P_v \| Q_{v,k}) \mid g], \\
e_{v,k}(g) &= D_{\mathrm{KL}}(Q_{v,k} \| S_{v,k}).
\end{aligned}
$$

The conditional projection identity is $t_{v,k}(g) = d_{v,k}(g) + e_{v,k}(g)$. We use 16,384 scrambled Sobol normal draws to form $\widehat{Q}_k$ and an independently scrambled 16,384 to estimate these conditional components. Their averages over the fixed held-out roots are denoted $\bar{t}_k^{\mathrm{obs}}, \bar{d}_k^{\mathrm{obs}}, \bar{e}_k^{\mathrm{obs}}$. They do not integrate over law-resampled Z_k and are not automatically the population terms $T_k, \Delta_k, \mathcal{E}_k$. We also compute a same-sample plug-in identity as a numerical invariant. Its maximum error is 3.55×10^{-15} over best-law contexts; the substantively relevant independent cross-fit error is at most 2.53×10^{-4} nats over every validation-competitive-law context and 2.35×10^{-4} over best-law contexts. Reconstructing every observed teacher logit from the sparse APPNP root weights has maximum absolute error 4.85×10^{-6}. The finite-matrix test compares sequential and union conditioning *with the same ridge* and therefore does not test the noiseless limit. Additional tests recover observed coordinates, permit zero variance when all target coordinates are observed, and check repeated and rank-deficient observations without injecting random variance.

Summary-matched oracle comparison. Let $R_k = [H_v, (MH)_v, \ldots, (M^k H)_v]$, $F = \sigma(A_{\mathrm{obs}}, R_k)$, and $G = \sigma(A_{\mathrm{obs}}, H_{B_k(v)})$. The latter refines the former because each hop summary is determined by the corresponding full local observation. With $Q_{\mathrm{sum}} = \mathbb{E}[P \mid F]$ and $Q_{\mathrm{ball}} = \mathbb{E}[P \mid G]$, conditional projection gives

$$
\mathbb{E}[D_{\mathrm{KL}}(P \| S) \mid F] = \mathbb{E}[d_{\mathrm{ball}} \mid F] + \mathbb{E}[D_{\mathrm{KL}}(Q_{\mathrm{ball}} \| Q_{\mathrm{sum}}) \mid F] + D_{\mathrm{KL}}(Q_{\mathrm{sum}} \| S).
$$

The middle term is nonnegative only after averaging the full-ball evidence conditional on the summary; subtracting two fixed-context residuals would not establish it. For each of four preselected test roots per seed, the Gaussian law supplies conditional target moments under both observations. We draw 64 outer full-ball refinements and 1,024 inner conditional teacher completions; the summary projection uses 4,096 scrambled Sobol draws. We reconstruct the trained student from its saved training seed and protocol, verifying its best epoch and realized teacher–student residual; no learner is refitted to the corresponding oracle. At $k = 0$, summary and ball coincide; the small numerical compression is within sampling discrepancy. Table 7 indicates small summary compression in these selected contexts, but the remaining oracle–student excess also includes possible access to A_{obs}, not only capacity or optimization.

Table 4: **Coordinate-conditioned observed-context decomposition.** Entries average conditional KL components over 32 fixed held-out roots for each of three trained teachers; Total is $\bar{t}_k^{\mathrm{obs}}$ and info. share is $\bar{d}_k^{\mathrm{obs}}/\bar{t}_k^{\mathrm{obs}}$. They are not population expectations over law-resampled Z_k. $[L, U]$ spans validation-competitive laws and seeds, not graph-population uncertainty.

Data	k	Total	$\bar{d}_k^{\mathrm{obs}}$	$\bar{e}_k^{\mathrm{obs}}$	info. share	law range $[L, U]$
Cora	0	0.3005	0.0069	0.2937	0.02	[0.0058,0.0081]
	1	0.0796	0.0013	0.0784	0.02	$[8.59 \times 10^{-4}$,0.0018]
	2	0.0269	2.57×10^{-4}	0.0266	0.01	$[1.47 \times 10^{-4}, 4.71 \times 10^{-4}]$
	3	0.0145	3.31×10^{-5}	0.0145	0.00	$[2.75 \times 10^{-5}, 3.84 \times 10^{-5}]$
	4	0.0099	4.23×10^{-6}	0.0099	0.00	$[3.42 \times 10^{-6}, 5.27 \times 10^{-6}]$
PubMed	0	0.2141	0.0037	0.2104	0.02	[0.0032,0.0043]
	1	0.0437	0.0018	0.0419	0.04	[0.0016,0.0020]
	2	0.0124	1.74×10^{-4}	0.0122	0.01	$[1.40 \times 10^{-4}, 2.08 \times 10^{-4}]$
	3	0.0049	3.31×10^{-5}	0.0049	0.01	$[2.69 \times 10^{-5}, 4.37 \times 10^{-5}]$
	4	0.0019	2.68×10^{-6}	0.0019	0.00	$[2.14 \times 10^{-6}, 3.29 \times 10^{-6}]$
Roman-empire	0	0.4230	0.1003	0.3227	0.24	[0.0832,0.1152]
	1	0.1440	0.0322	0.1119	0.22	[0.0313,0.0355]
	2	0.0701	0.0073	0.0628	0.10	[0.0052,0.0093]
	3	0.0446	0.0011	0.0435	0.02	$[8.78 \times 10^{-4}$,0.0013]
	4	0.0401	1.65×10^{-4}	0.0399	0.00	$[1.42 \times 10^{-4}, 2.12 \times 10^{-4}]$

Cora best γ: 1,1,1 (scale 0.61,0.60,0.60); PubMed best γ: 0.75,0.75,0.75 (scale 0.84,0.84,0.85); Roman-empire best γ: 0,0,0 (scale 0.59,0.59,0.59). Maximum independent closure error: 2.35×10^{-4} nats.

Table 5: **Conditioning-set audit.** Fit-only conditions on retained artifacts; train-aware additionally conditions on the realized training-side logit field. The latter is primary.

Data	Conditioning	k	Total	$\bar{d}_k^{\mathrm{obs}}$	info. share
Cora	Fit-only	0	0.8052	0.1006	0.12
		2	0.0636	0.0020	0.03
		4	0.0098	2.29×10^{-5}	0.00
	Train-aware	0	0.3005	0.0069	0.02
		2	0.0269	2.57×10^{-4}	0.01
		4	0.0099	4.23×10^{-6}	0.00
PubMed	Fit-only	0	0.2424	0.0433	0.18
		2	0.0588	8.93×10^{-4}	0.02
		4	0.0025	7.99×10^{-6}	0.00
	Train-aware	0	0.2141	0.0037	0.02
		2	0.0124	1.74×10^{-4}	0.01
		4	0.0019	2.68×10^{-6}	0.00
Roman-empire	Fit-only	0	0.3273	0.2648	0.81
		2	0.0700	0.0194	0.28
		4	0.0399	3.93×10^{-4}	0.01
	Train-aware	0	0.4230	0.1003	0.24
		2	0.0701	0.0073	0.10
		4	0.0401	1.65×10^{-4}	0.00

Table 6: **Observed-context same-information capacity intervention.** Both students use the same local observation and fixed test contexts, so $\bar{d}_k^{\mathrm{obs}}$ is unchanged; student excess is relative to the declared oracle.

Data	k	$\bar{e}_k^{\mathrm{obs}}$ (64)	$\bar{e}_k^{\mathrm{obs}}$ (256)	reduction
Cora	0	0.2937	0.3114	-6.0%
	2	0.0266	0.0247	7.3%
	4	0.0099	0.0078	20.5%
PubMed	0	0.2104	0.2104	0.0%
	2	0.0122	0.0104	15.0%
	4	0.0019	0.0015	23.9%
Roman-empire	0	0.3227	0.3297	-2.2%
	2	0.0628	0.0440	29.9%
	4	0.0399	0.0196	50.9%

Table 7: **Nested summary/ball oracle diagnostic under the same A_{obs} and frozen Gaussian law.** Twelve fixed test contexts per graph and radius (four preselected roots per seed) use the trained width-64 students. The columns implement $\mathbb{E}[\mathrm{KL}(P\|S) \mid F] = \mathbb{E}[d_{\mathrm{ball}} \mid F] + \mathbb{E}[\mathrm{KL}(Q_{\mathrm{ball}}\|Q_{\mathrm{sum}}) \mid F] + \mathrm{KL}(Q_{\mathrm{sum}}\|S)$, where $F = \sigma(A_{\mathrm{obs}}, R_k)$ and the full ball is resampled conditional on F. Values are context means; closure is the maximum absolute Monte Carlo discrepancy among the twelve contexts. The remaining excess includes any oracle information-access mismatch and is not pure optimization error.

Data	k	Total	Ball info.	Compression	Oracle–student excess	Max closure
Cora	0	0.5679	0.005125	$< 10^{-8}$	0.5628	$2.70 \cdot 10^{-5}$
	2	0.0260	0.000135	$7.41 \cdot 10^{-5}$	0.0258	$1.80 \cdot 10^{-5}$
	4	0.00993	$3.75 \cdot 10^{-6}$	$1.00 \cdot 10^{-5}$	0.00991	$1.40 \cdot 10^{-6}$
PubMed	0	0.1424	0.003668	$< 10^{-8}$	0.1388	$2.08 \cdot 10^{-5}$
	2	0.0193	0.000238	$9.20 \cdot 10^{-5}$	0.0190	$9.09 \cdot 10^{-6}$
	4	0.00242	$2.24 \cdot 10^{-6}$	$1.73 \cdot 10^{-5}$	0.00240	$2.60 \cdot 10^{-6}$
Roman-empire	0	0.4182	0.09990	$4.96 \cdot 10^{-6}$	0.3182	0.00139
	2	0.0715	0.003413	0.001215	0.0668	0.000245
	4	0.0412	0.000111	$8.50 \cdot 10^{-5}$	0.0410	$1.11 \cdot 10^{-5}$

Full training-observation span sensitivity. Let G denote the ten-step APPNP propagation operator and set $K = 4$, the largest student radius in the post-training observed-context audit. We enlarge the declared side information to

$$A_{\text{full}} = \sigma\big(A_{\text{fit}},\ H_{\mathcal{V}_{\text{law}}}, \{(M^r H)_u : u \in \mathcal{T} \cup \mathcal{V}_{\text{es}},\ 0 \le r \le K\},\ \{(GH)_u : u \in \mathcal{T} \cup \mathcal{V}_{\text{es}}\}\big). \tag{59}$$

Using the same maximum-radius information for every k keeps the side-information sigma-field fixed, so changes across radius still isolate local observations. The hop summaries are the student inputs, and GH gives the class-centered teacher targets used for distillation and early stopping.

Under the fitted matrix-normal law $H = \mathbf{1}\mu^\top + B_\gamma E$, all variables in Eq. (59) are linear observations LH. For APPNP root weights g_v, separability gives $\mathrm{Cov}(g_v^\top H \mid A) = \nu_v(A) t_\gamma \Sigma$. We report the relative scalar factor

$$\rho_v(A) = \frac{\nu_v(A)}{\nu_v(\varnothing)}. \tag{60}$$

The numerator factor for $A = A_{\text{full}}$ is the squared distance from $B_\gamma^\top g_v$ to the row span of LB_γ. We approximate this distance by matrix-free LSQR without materializing a node covariance. Any feasible least-squares iterate gives a *numerical upper bound* on the exact conditional variance. The audit uses eight prespecified roots for each of three seeds and the selected fitted law per dataset. We rerun all eight largest residuals for up to 6,000 LSMR iterations and retain the smaller feasible residual. One PubMed upper bound decreases from 0.571 to 0.507; the other seven change by at most 1.4×10^{-5}. Small bounds certify low variance, whereas large bounds do not certify positive residual variance. Because $\sigma(A_{\text{full}}, Z_k)$ refines A_{full}, Theorem 3 also gives the rootwise upper bound

$$\Delta_{v,k}(A_{\text{full}}) \le \tfrac{1}{4} t_\gamma \operatorname{tr}(\Sigma) \nu_v(A_{\text{full}}), \qquad 0 \le k \le K. \tag{61}$$

This bound relates the span audit to the information limit but does not identify either KL component under A_{full}.

Table 8: **Full training-observation span sensitivity.** Relative pre-local teacher-logit variance under coordinate-conditioned A_{obs} and the enlarged, radius-invariant A_{full} (24 contexts per dataset). The A_{full} columns are finite-solver upper bounds, not certified positive residuals; the count reports bounds below 10^{-4}. This is not a KL decomposition.

Data	med. $\rho(A_{\text{obs}})$	med. $\rho(A_{\text{full}})$	90th pct.	count	max.
Cora	0.367	$< 10^{-10}$	$< 10^{-10}$	22/24	0.398
PubMed	0.663	$< 10^{-6}$	0.484	18/24	0.730
Roman-empire	0.731	$< 10^{-10}$	$< 10^{-10}$	24/24	$< 10^{-10}$

The computed upper bounds fall below 10^{-4} of prior variance in 64 of 72 contexts, including all Roman-empire contexts. For two Cora and six PubMed contexts, even the extended numerical upper bounds exceed 10^{-2}; true variance there may be smaller because the linear inverse can be ill-conditioned. The stronger information set thus certifies near-determination in many contexts, but the remaining eight are unresolved by this audit. It does not replace the coordinate-conditioned A_{obs} KL decomposition.

Numerical budget. The randomized-Sobol audit uses eight roots per seed at radii 0, 1, and 4, with three scrambles at 512–16,384 points. Its between-scramble SD measures integration variability, not conditioning-system bias. The noiseless solver audit freezes the best Gaussian law, teachers, roots, and splits; for all 96 fixed roots per graph and every $k = 0, \dots, 4$, it compares 10^{-6}, 10^{-8}, 10^{-10} and a stable zero-ridge solution. Large systems use factor-based preconditioned conjugate gradients, with residuals reported separately. Rank-deficient small systems use an eigenspace pseudoinverse with tolerance $10^{-11} \max(\lambda_{\max}, 1)$; only negative roundoff variance within $10^{-9} \max(\nu_{\text{prior}}, 1)$ is clipped to zero, and a larger violation aborts. The 10^{-6} floor is active in 17, 54, and 3 of 96 radius-four contexts, respectively. The largest matched-QMC changes in d are below $2 \cdot 10^{-6}$ nats, so the coarse diagnosis is unchanged; this does not license many significant digits for near-zero high-radius gaps.

The interval in Table 4 takes the range over all validation-competitive candidates for each seed and then the union across seeds. It is a finite-candidate sensitivity range, not a confidence interval

Table 9: **Noiseless conditional-Gaussian solver audit.** Frozen best Gaussian laws, three seeds, 32 fixed test roots, and all five radii give 480 contexts per dataset. The first column counts radius-four contexts at the 10^{-6} variance floor (out of 96). Maximum $|d|$ changes compare the no-ridge, zero-floor reference with each ridge using identical 4,096-point scrambled Sobol draws; they are numerical sensitivity, not a confidence interval. The last column is the median between-scramble SD for d_4 at 16,384 points, which measures a different error source.

Data	k=4 floor	max $\lvert\delta d\rvert$: 10^{-6}	10^{-8}	10^{-10}	Solve resid.	QMC SD, k=4
Cora	17/96	$1.00 \cdot 10^{-6}$	$1.00 \cdot 10^{-8}$	$1.00 \cdot 10^{-10}$	$7.22 \cdot 10^{-16}$	$5.46 \cdot 10^{-12}$
PubMed	54/96	$3.73 \cdot 10^{-7}$	$3.73 \cdot 10^{-9}$	$3.73 \cdot 10^{-11}$	$9.49 \cdot 10^{-13}$	$2.16 \cdot 10^{-10}$
Roman-empire	3/96	$1.99 \cdot 10^{-6}$	$1.99 \cdot 10^{-8}$	$1.99 \cdot 10^{-10}$	$< 10^{-15}$	$4.07 \cdot 10^{-9}$

for the unknown true-law quantity. Reproducibility records include context rows, law diagnostics, summaries, intervals, software versions, and deterministic seeds.

C.2 Real-topology conditional-completion population

For each dataset, the topology and official split 0 are fixed. A two-layer feature MLP with hidden width 64, dropout 0.35, AdamW learning rate 0.01, weight decay 5×10^{-4}, maximum 180 epochs, and validation patience 25 is trained on labels. We class-center its logits and estimate μ and Σ using training nodes only; a 10^{-6} ridge is added on the class-centered subspace.

The population unit is a test-root and base-logit world on that fixed topology. Conditional on the calibrated parameters, node logits are i.i.d. $H_u \sim \mathcal{N}(\mu, \Sigma)$, and the root is uniform over the test nodes. We sample 32 roots without replacement. A fresh anchor world is drawn for each root and teleport probability. At radius k, its logits on $B_k(v)$ are observed and held fixed; all remote logits follow the independent Gaussian conditional law. The same anchor is used across radii, so the observations are nested.

Because PPR is linear, no full remote graph world needs to be materialized. Writing $g_u = (G_\alpha)_{vu}$, the conditional root logit is Gaussian with

$$m_{v,k} = \sum_{u \in B_k(v)} g_u H_u + \sum_{u \notin B_k(v)} g_u \mu, \tag{62}$$

$$\Sigma_{v,k} = \left(\sum_{u \notin B_k(v)} g_u^2 \right) \Sigma. \tag{63}$$

We draw 16,384 samples from this distribution to estimate Q_k, and an independent 16,384 to evaluate KL. The exact covariance above supplies the logit upper bound; the evaluation batch supplies the probability-variance lower estimate and candidate KL. For each anchor and root, this procedure estimates the conditional quantity $d_{v,k}$. Averaging over anchors drawn from the declared i.i.d. population estimates Δ_k. Differences of these population averages estimate shell values; the first threshold crossings of the context estimates define $\widehat{\tau}_{\varepsilon,8}$. We do not infer pathwise monotonicity from the monotonicity of Δ_k. Random seeds are deterministic functions of dataset, teleport probability, root, and radius. The audit records context-level estimates, PPR residuals, root-cluster summaries, software versions, entropy plug-in estimates, Fisher–Green terms, interval gains, and stopping decisions.

C.3 Completion-law sensitivity on fixed topologies

The sensitivity study keeps the topology, feature encoder, training-calibrated class covariance, roots, PPR teachers, radii, and Monte Carlo sample counts fixed. It replaces i.i.d. node logits by a matrix-normal Gaussian law with node correlation

$$K_\gamma = D_\gamma^{-1/2} B_\gamma^2 D_\gamma^{-1/2}, \qquad B_\gamma = I + \gamma M, \qquad D_\gamma = \operatorname{diag}(B_\gamma^2).$$

The standardization gives $(K_\gamma)_{uu} = 1$, so each node retains marginal class covariance Σ. The cases $\gamma = 0, 0.2, 0.5$ define the i.i.d. and two graph-correlated laws. The main sensitivity analysis draws

one anchor from the primary i.i.d. law and reuses its observed H_L under all three laws. Every nonsingular Gaussian law assigns positive density to that local value, so each conditional distribution is well defined. This design holds the realized local evidence fixed rather than merely coupling different evidence through common noise.

For local set $L = B_k(v)$, remote set R, and remote PPR weights g_R, let $c = K_{\gamma,LR} g_R$. Gaussian conditioning gives

$$\mathbb{E}[g_R^\top H_R \mid H_L] = (g_R^\top \mathbf{1})\mu + c^\top K_{\gamma,LL}^{-1}(H_L - \mathbf{1}\mu^\top), \tag{64}$$

$$\mathrm{Cov}(g_R^\top H_R \mid H_L) = \left(g_R^\top K_{\gamma,RR} g_R - c^\top K_{\gamma,LL}^{-1} c\right)\Sigma. \tag{65}$$

The implementation applies K_γ through two sparse multiplications and solves the local system by conjugate gradients. It never materializes the dense node covariance. All 3,840 contexts satisfy relative PPR and linear solve tolerances of 10^{-11} and 10^{-10}, respectively. The 1,280 i.i.d. rows reproduce the primary run: every reported KL statistic differs by at most 1.3×10^{-15}, and exact covariance quantities by at most 1.1×10^{-14}.

For $\gamma > 0$, averaging $d_k^{(\gamma)}$ over these i.i.d. anchors estimates $\bar{d}_k^{(\gamma|\mathrm{i.i.d.})}$, not the native population $\Delta_k^{(\gamma)}$. We therefore call differences across measured radii "residual reductions" rather than conditional mutual information. A companion analysis samples a standardized filtered anchor from each law; it identifies the three law-native population profiles. Under that analysis, $\gamma = 0.5$ changes the mean radius-zero Δ_k by 15–64% across datasets. The two protocols answer different questions and agree that the diagnosis is law-sensitive.

Table 10: Context-wise Fisher–Green accuracy under matched local observations. Relative error is $|\frac{1}{2}\,\mathrm{tr}(J_{v,k}\Gamma_{v,k}) - \widehat{d}_{v,k}|/\widehat{d}_{v,k}$.

Completion law	Contexts	Spearman correlation	Median relative error
i.i.d.	1,280	0.9999	0.82%
Graph-correlated, $\gamma = 0.2$	1,280	0.9999	0.85%
Graph-correlated, $\gamma = 0.5$	1,280	0.9999	0.83%

C.4 Specified non-Gaussian completion sensitivity

The Gaussian conclusions are checked against a covariance-matched global Student-t_5 node-logit field. Let $r = C - 1$, $G \sim \chi^2_\nu$, and let Z_u be independent centered Gaussian vectors on $\mathbf{1}^\perp$ with covariance $\Sigma_s = (\nu - 2)\Sigma/\nu$. We define

$$H_u = \mu + \sqrt{\nu/G}\, Z_u, \qquad \nu = 5.$$

Thus each node has mean μ and covariance Σ; distinct nodes are uncorrelated but remain dependent through the shared radial variable. For $L = B_k(v)$, set $d_L = r|L|$ and $\delta_L = \sum_{u\in L}(H_u - \mu)^\top \Sigma_s^+ (H_u - \mu)$. Conditional on H_L, the remote PPR aggregate is Student-$t_{\nu+d_L}$ with scale

$$\frac{\nu + \delta_L}{\nu + d_L}\, \|g_R\|^2_{2\Sigma_s}.$$

This provides analytic conditional sampling without changing the topology, teacher, roots, radii, marginal covariance, or Monte Carlo budget. We reuse the anchor drawn from the i.i.d. Gaussian law under both laws, so the comparison concerns matched conditional residuals rather than the Student-t law's native Δ_k.

Across 2,560 contexts, the Student-t_5 law changes mean radius-zero residuals by -3.3% to 7.9% relative to Gaussian. Mean $\widehat{\tau}_{10^{-4},8}$ is unchanged on three datasets and changes by 0.031 hops on PubMed. The overall Fisher–Green Spearman correlation is 0.9998, with 0.84% median relative error, compared with 0.9999 and 0.82% under Gaussian. Heavy-tailed higher-order structure therefore changes residual magnitude without changing the qualitative radius profile or the second-order mechanism diagnosis in this audit.

Table 11: Observed-context radius-zero information-share ranges over jointly NLL-competitive candidate laws and three teacher seeds; shares are ratios of conditional components averaged over matched fixed roots. These are finite-candidate sensitivity ranges, not confidence intervals. Roman-empire crosses the 50% diagnostic threshold only when the declared conditioning set changes.

Dataset	Pairs	A_{fit}	A_{obs}
Cora (Gaussian only)	3	11.8–13.6%	1.9–3.4%
PubMed	36	5.0–22.6%	1.3–2.9%
Roman-empire	36	61.9–83.9%	20.5–30.8%

C.5 Cross-family fitted-law predictive and sensitivity audit

We additionally screen covariance-matched, graph-correlated global elliptical Student-t fields against the fitted Gaussian family. This is separate from the standardized i.i.d.-anchor audit above. For $\nu \in \{4, 5, 8, \infty\}$, let

$$H = \mathbf{1}\mu^\top + B_\gamma E, \qquad E = \sqrt{\frac{\nu-2}{G}}\, Z, \quad G \sim \chi^2_\nu, \quad Z_{u,:} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, t_{\gamma,\nu}\Sigma), \tag{66}$$

with the Gaussian limit used at $\nu = \infty$. One scale mixture is shared across the field, so at a fixed temperature $t_{\gamma,\nu}$ the law has the same nodewise covariance and graph kernel as its Gaussian counterpart, but different joint tails. Temperatures are then calibrated separately across candidates. We keep the trained teacher, student, graph, root set, A_{fit}, A_{obs}, and Z_k fixed.

For d observed class-contrast coordinates and unscaled Mahalanobis value δ, Gaussian conditioning gives the same conditional mean and Schur covariance C as before. The conditional Student-t degrees of freedom are $\nu + d$, and its scale matrix is

$$\frac{\nu - 2 + \delta/t_{\gamma,\nu}}{\nu + d}\, t_{\gamma,\nu} C. \tag{67}$$

We minimize conditional predictive NLL on the same law-selection nodes for each (γ, ν), calibrating the scalar temperature in $[0.1, 4]$. The full γ prescreen is $\{0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2, 4, 8, 16, 32\}$ on every dataset. It confirms that all retained candidates lie within the conditional-KL evaluation windows $[0.35, 1.5]$ on PubMed and $[0, 0.5]$ on Roman-empire; Cora uses the full grid throughout. Candidates within 0.01 NLL per contrast dimension of the selection winner form the finite validation-competitive set. The untouched audit nodes then report conditional NLL and 50/80/95% joint Mahalanobis coverage; they do not affect selection. Any interval over the retained candidates is a *finite-family sensitivity envelope*, not a confidence interval for an unknown true law. Because the high-dimensional anchor field makes $\nu + d$ large, the fitted heavy-tailed alternatives may fail to add a distinct validation-competitive model; that outcome is reported rather than relabelled as robustness. Even a retained finite-ν model can be nearly Gaussian after conditioning on thousands of coordinates; competition within this global-scale-mixture family is therefore narrower than robustness to arbitrary non-Gaussian completions. Test-root KL components under A_{obs} use independent 4,096-point scrambled Sobol batches, one for Q_k and one for evaluation. The finite-ν radius-zero A_{fit} sensitivity uses 65,536 points in each independent batch on PubMed and Roman-empire; Gaussian values reuse the 16,384-point audit. Their rootwise average is an observed-context conditional audit, not an expectation over law-resampled local observations.

We therefore run a separate law-native radius-zero information audit. For the best Gaussian and best finite-ν candidate at each seed, we draw 32 new root-logit contexts from the candidate conditional law given fixed A_{obs} at each of four matched test roots. For every redrawn Z_0, 4,096 conditional completions fit Q_0 and 4,096 independent completions evaluate its KL loss. The resulting average estimates $\mathbb{E}_{Z_0|A_{\mathrm{obs}},L}[d_0(Z_0)]$ over the selected roots, conditional on the fixed observed A_{obs} and fitted law L. This is a local-context population average for the declared post-training conditional population, not an unconditional expectation over training-side fields or a graph-wide root population. It does not report a student information share, because the finite-student comparison in the main audit is deliberately tied to observed test-root contexts. We identify whether the best finite-ν candidate passed the selection criterion and never treat a rejected candidate as part of the sensitivity envelope.

The best Cora Student-t candidates miss the primary NLL tolerance by 0.0169–0.0215 per contrast dimension, although their untouched 95% coverage is closer to nominal (94.5–95.5% versus 87.9–

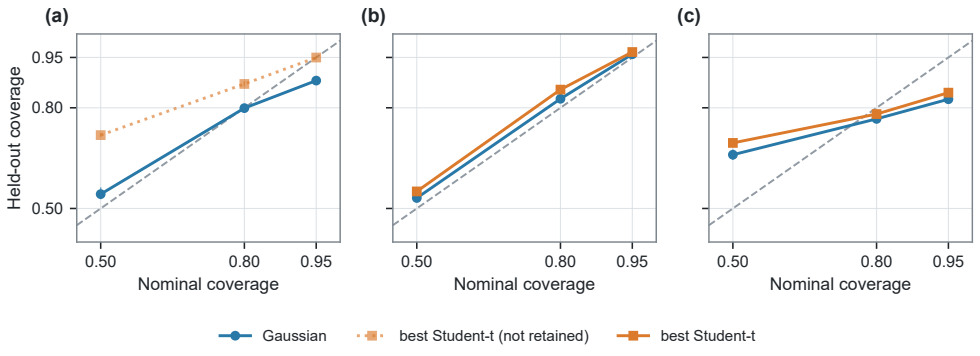


Figure 5: **Untouched-node predictive coverage across fitted law families.** (a) Cora, (b) PubMed, and (c) Roman-empire. Curves average three teacher seeds; thin vertical segments span their seedwise values, and the gray diagonal is nominal coverage. The best finite-ν candidate is shown for each seed even when it misses the prespecified NLL qualification threshold; dotted orange on Cora is therefore excluded from Table 11. Coverage is descriptive across dependent nodes in one graph. In particular, neither retained family adequately covers Roman-empire's upper tail.

88.4% for Gaussian). We therefore show them in Figure 5 but exclude them from the primary sensitivity envelope. On Roman-empire, even the best finite-ν candidates cover only 83.6–85.9% at a nominal 95%; this failure remains a substantive limitation of the fitted family. An exploratory tolerance sensitivity widens the NLL cutoff to 0.025 and adds Cora finite-ν and adjacent Gaussian candidates. Its information shares still remain below 50% under both A_{fit} (8–18%) and A_{obs} (1.7–4.3%). This exploratory check does not change the prespecified retained set; its heavy-tailed A_{fit} tail estimates have coarser QMC closure and support only the qualitative diagnosis. For the primary retained set, the increased finite-ν A_{fit} budget reduces the largest Roman-empire independent-batch closure discrepancy to 1.50×10^{-3} nats, 0.11% of that context's total KL. Across datasets and conditioning regimes, the 95th percentile absolute closure discrepancy is below 10^{-4} nats. This is a numerical audit of direct component estimates, not a claim that finite-sample closure is algebraically exact.

Table 12: Conditional law-native radius-zero information residual in nats after resampling local $Z_0 \mid A_{\text{obs}}$ at four selected roots (range over three teacher seeds; 128 outer contexts per seed). The Student-t column is an exploratory stress test on Cora because no finite-ν candidate passed the primary selection rule there. Unlike Table 11, this integrates over law-redrawn local observations and does not include a finite-student share.

Dataset	Gaussian	Best Student-t
Cora	0.00096–0.00540	0.00146–0.00705
PubMed	0.00213–0.00978	0.00223–0.01065
Roman-empire	0.0832–0.211	0.0874–0.225

C.6 Nonlinear conditional-KL decomposition

To test whether empirical KL accounting depends on APPNP linearity, we replace only the frozen teacher map by the contractive equilibrium

$$X_\infty = H + q\tanh(MX_\infty), \qquad q \in \{0.4, 0.7\}, \quad P_v = \text{softmax}(X_{\infty,v}). \tag{68}$$

The centered logit field H, fitted Gaussian law, A_{obs}, radius-k observation, and root protocol are unchanged. Students retain the same width-64 MLP architecture, hop-summary inputs, optimizer, split, and early-stopping protocol, but must be retrained against each changed teacher target. This is a controlled map intervention on trained base logits, not a separately optimized nonlinear-GNN benchmark. Cora and Roman-empire use three model seeds, the first four roots of the 32-root APPNP audit for each seed, and $k \in \{0, 2, 4\}$.

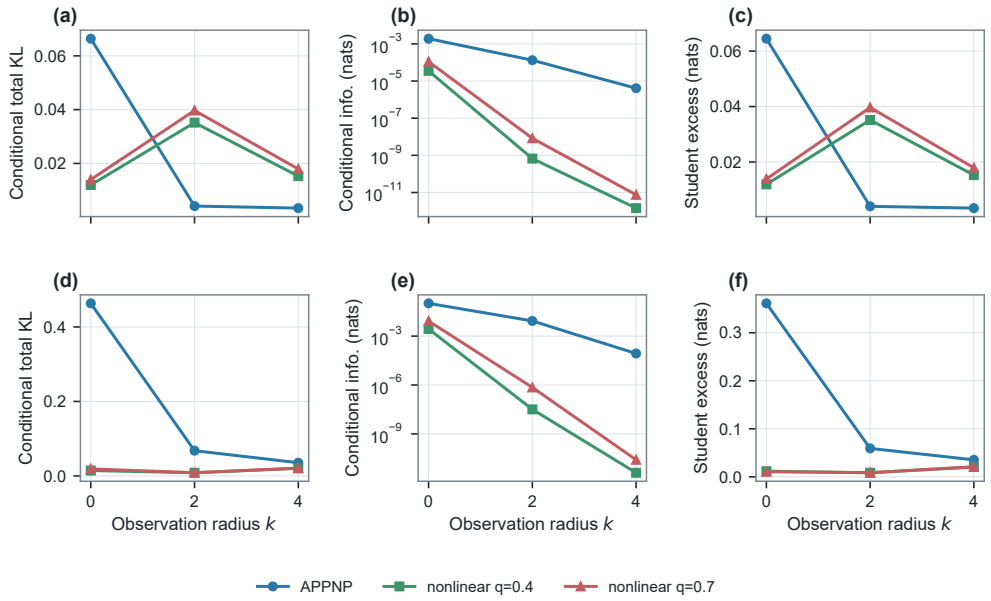


Figure 6: **Direct observed-context accounting with a contractive nonlinear teacher.** (a–c) Cora total, information (log scale), and realization KL; (d–f) the corresponding Roman-empire quantities. Curves average three teacher seeds and four matched roots per seed at each radius. The APPNP reference uses its 16,384-point independent QMC audit; the nonlinear map uses 4,096 conditional samples for Q_k and 4,096 independent evaluation samples. Changing the teacher also requires refitting the student under the unchanged learner protocol; the curves are not an architecture ranking. Extremely small high-radius information values approach numerical resolution and should not be read as precisely identified magnitudes.

For each fixed observed (A_{obs}, Z_k), Gaussian Matheron correction generates independent compatible logit fields for projection fitting and evaluation. Each completion is propagated through the nonlinear fixed point. We compute Q_k from 4,096 fit completions, and compute total KL and information KL directly from 4,096 independent evaluation completions. The realization KL is evaluated directly from the fitted Q_k and the frozen student. No KL component is constructed by subtraction. For each q, we choose the smallest iteration count $t_q \le 100$ satisfying $\sqrt{nC}\, q^{t_q}/(1-q) \le 10^{-6}$. The root iterate depends only on the radius-(t_q-1) logit ball. Starting at $X_0 = 0$, the first difference after $X_1 = H$ is bounded by $q\sqrt{nC}$ because $|\tanh| \le 1$. Since $\|M\|_2 \le 1$ and $\tanh$ is 1-Lipschitz, summing subsequent contractive differences gives this conservative full-graph Frobenius bound on its remaining root-logit Euclidean error. We also record the realized fixed-point residual, conditional-anchor error, and independent-batch decomposition closure. Values below Monte Carlo resolution are reported qualitatively.

Across the 144 nonlinear context rows, the 95th percentile absolute independent-batch closure error is 6.97×10^{-5} nats and the maximum is 1.40×10^{-3} nats. The largest fit/evaluation discrepancy in Q_k is 7.61×10^{-3} in ℓ_1 norm. By contrast, the certified fixed-point root-logit truncation bound is at most 9.71×10^{-7} and the largest realized fixed-point residual is 4.77×10^{-7}. The worst closure is therefore a finite conditional-sampling effect, not fixed-point truncation.

C.7 Formal conditional-completion population

The controlled population is defined directly over shell statistics rather than over sampled adjacency matrices. A graph world contains $X = (X_0, \ldots, X_5)$, where the six coordinates are independent standard Gaussian variables. The radius-k observation is $Z_k = (X_0, \ldots, X_k)$. For condition $d \in \{\mathrm{fast}, \mathrm{slow}, \mathrm{far}\}$, define

$$P_d(X) = \eta + (1 - 2\eta)\sigma(w_d^\top X), \qquad \eta = 0.02,$$

with

$$w_{\text{fast}} = (1.20, 0.70, 0.35, 0.15, 0.07, 0.03),$$
$$w_{\text{slow}} = (1.00, 0.75, 0.60, 0.45, 0.30, 0.20),$$
$$w_{\text{far}} = (0.70, 0.20, 0.10, 0.05, 0.60, 0.80).$$

Conditional on $Z_k = z_k$, a compatible completion keeps the first $k+1$ coordinates fixed and independently redraws every coordinate $X_r \sim \mathcal{N}(0,1)$ for $r > k$. This is exactly the conditional law of $X \mid Z_k = z_k$. We sample 384 contexts, use 4,096 completions to form the reference conditional mean, and use 2,048 independent completions to evaluate its KL risk. Cross-fitting uses independent fitting sizes $M \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$, 256 evaluation completions, and 30 repeats. The construction isolates conditional projection under a fully specified sampling law while keeping the graph topology fixed.

Finite-student decomposition. For each law and radius, a two-hidden-layer width-64 MLP receives only Z_k and is trained by soft-label binary cross-entropy against $P_d(X)$. Training sets contain 512, 2,048, or 8,192 independently sampled worlds; each setting uses five initialization and data seeds plus an independent validation set of 4,096 worlds. We evaluate Q_k by 64-point Gauss–Hermite quadrature, then use 8,192 new contexts and 64 independent conditional completions per context to estimate $T_k = \mathbb{E}D_{\mathrm{KL}}(P\|S_k)$, $\Delta_k = \mathbb{E}D_{\mathrm{KL}}(P\|Q_k)$, and $\mathcal{E}_k = \mathbb{E}D_{\mathrm{KL}}(Q_k\|S_k)$. Thus all three terms share the same teacher, law, observation, student, and held-out population; no term is inferred by subtracting the other two. The maximum absolute closure error at 8,192 training worlds is 3.19×10^{-5}.

Data and splits. Cora contains 2,708 nodes and 10,556 directed entries after symmetrization; PubMed contains 19,717 nodes and 88,648 entries; Roman-empire contains 22,662 nodes and 65,854 entries; Amazon-ratings contains 24,492 nodes and 186,100 entries. We use five of the ten official 60/20/20 masks for every dataset. Planetoid bag-of-word features are row-normalized. Heterophilous-graph features are standardized using training nodes only.

Propagation operator. Every graph is first converted to an undirected graph. We then add self-loops and use the symmetric GCN normalization $M = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$. Thus M is symmetric, nonnegative, and satisfies $\|M\|_2 \le 1$, matching the PPR, spectral, and nonlinear fixed-point assumptions. A radius-k observation contains the normalized weights on all used walks, including degrees required to normalize edges incident to boundary nodes.

Teacher and student architecture. Each teacher first projects inputs to width 64, then applies four residual message-passing blocks with GELU, LayerNorm, and dropout, followed by a linear classifier. The GATv2 blocks use four heads with averaged outputs. The hop-MLP concatenates $X, \widehat{A}X, \ldots, \widehat{A}^k X$ and uses two hidden layers. The local GCN has k residual graph-convolution blocks. At $k = 0$ both reduce to feature-only MLPs. The copied control has the teacher's architecture and parameters; at $k = 4$ it is identical by construction.

KRD adapter. Following the public implementation of Wu et al. (2023b), we measure each training node's teacher-entropy change after unit Gaussian feature noise, normalize it to $[0, 1]$, and sample training-edge teacher targets with probability $1 - d_i^p$. The exponent is updated every ten epochs after epoch 50 from binned teacher–student agreement with momentum 0.99. Self-pairs are always retained. To enforce the strict protocol, both endpoints are training nodes; no test-node teacher target enters the loss.

TINED adapter. Following Zhou et al. (2025), the student mirrors the sequence of feature transformations, initializes every shape-compatible affine transform and normalization from the teacher, and minimizes the squared error between teacher and student log Dirichlet-energy ratios on training-induced edges. We use coefficient 0.1. For GATv2, head-specific affine weights are averaged before injection. This common-backbone adapter preserves the mechanism but does not reproduce the authors' DGL architecture or dataset-specific search.

Table 13: **Exact global-PPR audit.** Values average four teleport probabilities, five official splits, and three seeds. Gain ranges are over teleport probabilities. The learned column reports the median radius attaining minimum hop-MLP KL in each block.

Dataset	KL k=0	KL k=16	Agree k=16	Teacher gain range	Learned best k
Cora	0.5202	0.0154	0.9776	[0.0865,0.1113]	8
PubMed	0.3115	0.0097	0.9850	[-0.0371,0.0222]	8
Roman-empire	1.0901	0.0537	0.9708	[-0.5431,-0.2691]	2
Amazon-ratings	0.1622	0.0097	0.9877	[-0.0010,0.0140]	8

Matched-control objectives. The hard-label weight is 0.5 and temperature is 1 for GLNN, KRD, and TINED adapters. Pure soft KD omits hard labels. The transductive GLNN variant uses all-node teacher targets exactly to expose how the information set changes; all other methods use teacher targets only on training nodes in the optimization loss. Every student is selected by validation teacher KL under the same 180-epoch and patience-25 budget, so validation targets are included in the declared training side information A.

Compute and reproducibility. Experiments ran with Python 3.13.5, PyTorch 2.13.0+cu130, PyTorch Geometric 2.8.0.post1, NumPy 2.4.4, SciPy 1.17.1, and pandas 3.0.3 on one NVIDIA RTX 4080. The 2,700-configuration paired local factorial required approximately 37 minutes across the recorded runs. Every row stores dataset, official split, model seed, teacher, student, radius, metrics, and early-stopping epoch. A bitwise protocol rerun agreed in accuracy and agreement and differed in KL by less than 10^{-6}.

The exact global-PPR matrix used 64-bit sparse propagation for fixed-point and analytic-tail computations, with relative tolerance 10^{-10} and a 2,000-iteration cap. All 240 teachers converged; the slowest setting required 299 iterations. The 2,880 method-level rows required 603.1 seconds on the same GPU. Raw metrics, teacher audits, run metadata, and theorem assertions are retained for reproducibility.

For each learned hop-MLP block b, we additionally compute $m_b = \max_k [D_{\mathrm{KL}}^{(b,k+1)} - D_{\mathrm{KL}}^{(b,k)}]_+$. Across all 240 blocks, the median, 90th percentile, and maximum of m_b are 6.46×10^{-4}, 5.18×10^{-2}, and 0.132. Upward excursions exceed 10^{-3} in 106 blocks, 10^{-2} in 59 blocks, and 5×10^{-2} in 27 blocks. These statistics use deterministic aggregation over all per-block audit values.

The nonlinear audit used the residual equilibrium $X = H + q \tanh(MX)$, a 10^{-10} relative fixed-point tolerance, and the same frozen base logits and data splits. All 240 teachers converged; the maximum residual was 7.83×10^{-11}. The 2,880 method-level rows required 543.3 seconds. The audit records per-node maximum certificate excess, test-subset aggregate-envelope excess, iteration count, teacher utility, and all student metrics. A second deterministic execution matched all 2,880 primary keys, reproduced accuracy and agreement exactly, and had maximum rowwise KL difference 3.00×10^{-19}. It is retained as a reproducibility audit rather than pooled into the statistical summaries.

D ADDITIONAL RESULTS

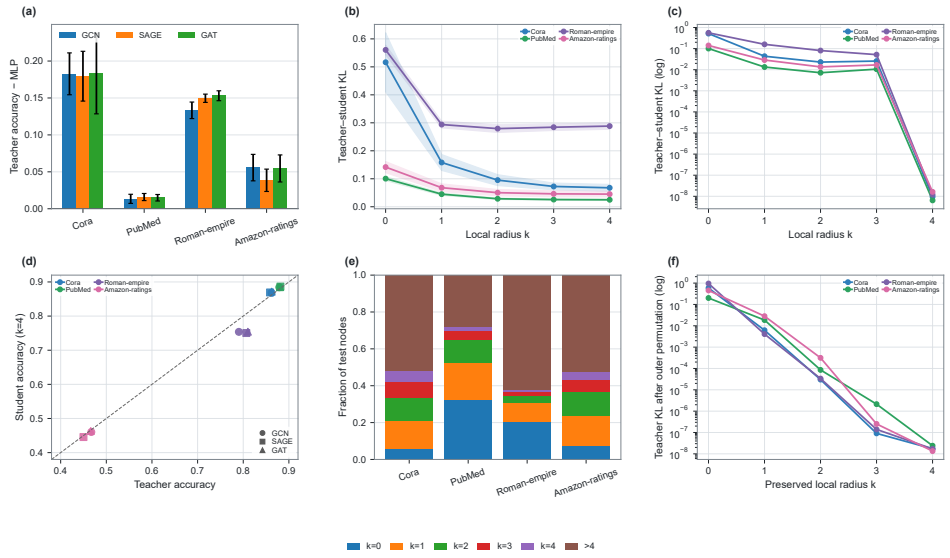


Figure 7: **Information versus finite realization.** Three finite-depth teacher families on four graphs separate teacher utility (a), hop-MLP fidelity (b), copied-parameter closure (c), task accuracy versus KL (d), law-aware diagnostic radii (e), and preserved-neighborhood sensitivity (f). These panels diagnose realization after fixing the information protocol.

Table 14: Nonlinear fixed-point audit. Coverage is the fraction of blocks that meet the persistent KL/agreement criterion at some measured radius; the median is conditional on coverage. Teacher-gain ranges span $q = 0.2$ to $q = 0.8$.

Dataset	Analytic coverage	Median (k)	Learned coverage	Teacher gain range
Cora	1.00	1.0	0.00	[0.0091, 0.0256]
PubMed	1.00	0.0	0.75	[0.0021, 0.0058]
Roman-empire	1.00	1.5	0.00	[-0.0386, -0.0017]
Amazon-ratings	1.00	2.0	0.00	[0.0038, 0.0067]

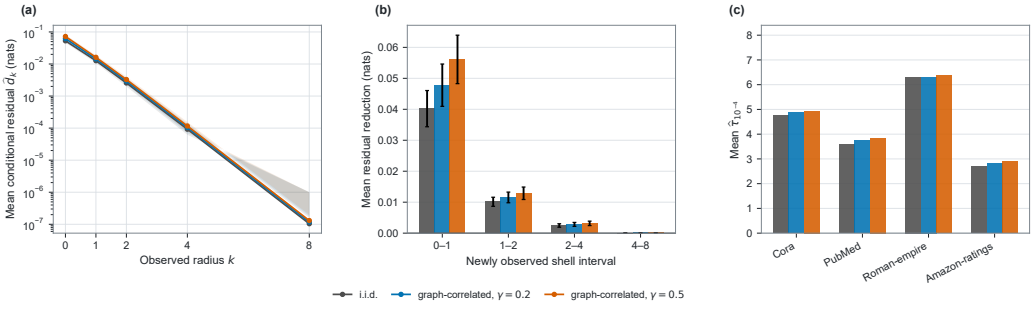


Figure 9: **Completion-law sensitivity on fixed real topologies.** (a) Matched-context residuals. (b) Paired reductions over measured intervals. (c) Mean $\widehat{\tau}_{10^{-4},8}$. Panels hold topology, marginals, roots, observations, teachers, radii, and budget fixed, varying only dependence.

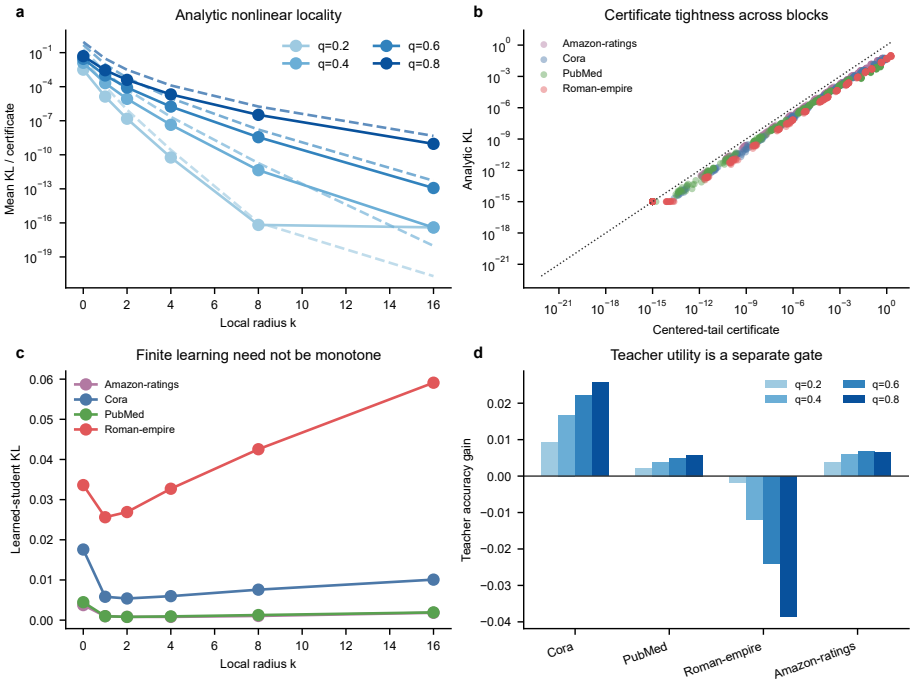


Figure 8: **Nonlinear global-diffusion audit.** Analytic locality and certificate tightness (a–b), finite-learning non-monotonicity (c), and the teacher-utility gate (d). The panels verify the nonlinear certificate while keeping it distinct from population identification.

Table 15: **Law-native completion sensitivity.** Plug-in population radius-zero residual and mean context stopping radius at $\varepsilon = 10^{-4}$, averaged over 32 roots and two teleport probabilities sampled under each law's own anchor distribution. These law-native estimands differ from the i.i.d.-reference matched-context quantities in Figure 9.

	Mean native $\widehat{\Delta}_0$ (nats)			Mean native $\widehat{\tau}_{10^{-4},8}$		
Dataset	i.i.d.	$\gamma = 0.2$	$\gamma = 0.5$	i.i.d.	$\gamma = 0.2$	$\gamma = 0.5$
Cora	0.05400	0.06054	0.06720	4.750	4.875	4.938
PubMed	0.02923	0.03164	0.03366	3.594	3.719	3.719
Roman-empire	0.12224	0.14883	0.17898	6.312	6.500	6.625
Amazon-ratings	0.00615	0.00782	0.01007	2.703	2.781	2.938

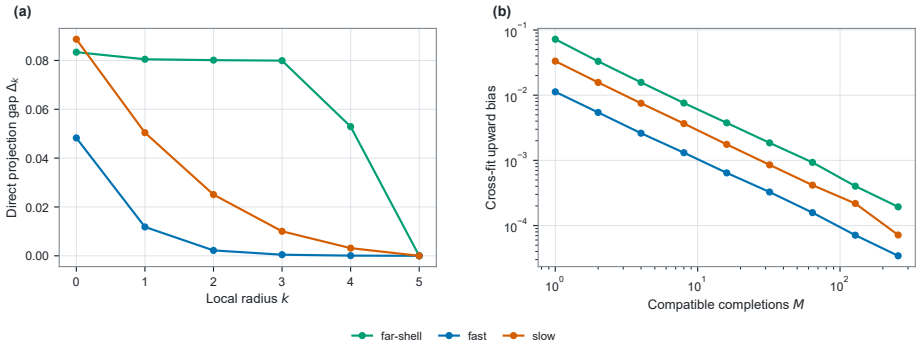


Figure 10: **Controlled conditional-completion validation.** Monte Carlo estimates of population gaps for fast, slow, and far-shell laws (left), and decay of the upward cross-fit bias with the number of compatible completions (right). This checks the estimator under a known population independently of PPR and real-topology calibration.

Table 16: **Non-Gaussian matched-context sensitivity.** The covariance-matched Student-t_5 field and Gaussian law share nodewise means and covariances. Averages use the same i.i.d.-Gaussian reference contexts and therefore are not law-native population values.

Dataset	Law	Mean $\widehat{d}_0$	Mean $\widehat{\tau}_{10^{-4},8}$	Fisher Spearman	Median rel. error
Cora	Gaussian	0.05400	4.750	0.9999	1.01%
Cora	Student-t_5	0.05751	4.750	0.9999	0.92%
PubMed	Gaussian	0.02923	3.594	0.9999	0.82%
PubMed	Student-t_5	0.02830	3.562	0.9996	1.04%
Roman-empire	Gaussian	0.12224	6.312	0.9994	1.42%
Roman-empire	Student-t_5	0.13196	6.312	0.9994	1.23%
Amazon-ratings	Gaussian	0.00615	2.703	1.0000	0.58%
Amazon-ratings	Student-t_5	0.00595	2.703	1.0000	0.56%

Table 17: **Exact empirical accounting for one finite student.** Slow law, width-64 MLP, and 8,192 training worlds; entries are means [seed s.d.] over five seeds. Δ_k uses the same held-out population; closure is maximum $|T_k - \Delta_k - \mathcal{E}_k|$.

k	Total T_k	Information Δ_k	Realization $\mathcal{E}_k$	Max closure error
0	0.089880 [0.000029]	0.089830	5.66×10^{-5} [2.79×10^{-5}]	1.46×10^{-5}
1	0.051797 [0.000164]	0.051657	1.37×10^{-4} [1.60×10^{-4}]	1.16×10^{-5}
2	0.025544 [0.000019]	0.025480	6.24×10^{-5} [1.66×10^{-5}]	5.39×10^{-6}
3	0.010204 [0.000009]	0.010149	5.67×10^{-5} [8.37×10^{-6}]	3.00×10^{-6}
4	0.003189 [0.000007]	0.003144	4.52×10^{-5} [6.62×10^{-6}]	1.09×10^{-6}
5	2.65×10^{-5} [3.01×10^{-6}]	0	2.65×10^{-5} [3.01×10^{-6}]	$< 4\times10^{-21}$

Table 18: **Finite-student sample-size audit.** Mean realization gap over five seeds for the same width-64 MLP, with Δ_k fixed by the declared population. Max closure error is taken over all three sample sizes and five seeds. The realization gap decreases as the number of training worlds increases from 512 to 2,048 to 8,192 in all 18 law–radius blocks.

Law	k	Δ_k	$\mathcal{E}_k$, 512	$\mathcal{E}_k$, 2,048	$\mathcal{E}_k$, 8,192	Max closure error
Fast	0	0.048693	3.47×10^{-4}	1.03×10^{-4}	2.27×10^{-5}	3.66×10^{-5}
Fast	1	0.011936	1.32×10^{-4}	5.67×10^{-5}	2.12×10^{-5}	1.01×10^{-5}
Fast	2	0.002263	8.38×10^{-5}	4.24×10^{-5}	2.21×10^{-5}	2.83×10^{-6}
Fast	3	0.000467	6.50×10^{-5}	3.05×10^{-5}	2.14×10^{-5}	1.02×10^{-6}
Fast	4	7.22×10^{-5}	3.63×10^{-5}	2.34×10^{-5}	2.01×10^{-5}	2.54×10^{-7}
Fast	5	0	3.69×10^{-5}	2.66×10^{-5}	2.39×10^{-5}	$< 7\times10^{-21}$
Slow	0	0.089830	2.69×10^{-4}	1.01×10^{-4}	5.66×10^{-5}	3.99×10^{-5}
Slow	1	0.051657	5.99×10^{-4}	1.85×10^{-4}	1.37×10^{-4}	1.16×10^{-5}
Slow	2	0.025480	4.72×10^{-4}	1.67×10^{-4}	6.24×10^{-5}	1.98×10^{-5}
Slow	3	0.010149	4.07×10^{-4}	1.63×10^{-4}	5.67×10^{-5}	1.02×10^{-5}
Slow	4	0.003144	2.23×10^{-4}	9.62×10^{-5}	4.52×10^{-5}	3.07×10^{-6}
Slow	5	0	3.61×10^{-5}	2.88×10^{-5}	2.65×10^{-5}	$< 7\times10^{-21}$
Far-shell	0	0.083036	2.93×10^{-4}	1.09×10^{-4}	5.05×10^{-5}	2.65×10^{-5}
Far-shell	1	0.080117	7.83×10^{-4}	2.86×10^{-4}	2.46×10^{-4}	5.67×10^{-5}
Far-shell	2	0.079290	1.23×10^{-3}	3.63×10^{-4}	1.44×10^{-4}	2.84×10^{-5}
Far-shell	3	0.079396	2.15×10^{-3}	5.29×10^{-4}	2.61×10^{-4}	4.86×10^{-5}
Far-shell	4	0.051864	1.42×10^{-3}	6.93×10^{-4}	2.42×10^{-4}	3.80×10^{-5}
Far-shell	5	0	2.42×10^{-5}	1.84×10^{-5}	1.53×10^{-5}	$< 4\times10^{-21}$

Table 19: Teacher–student KL (↓) for feature-only controls: means [95% split-cluster bootstrap intervals] over five splits. Strict rows use training-node targets; the final row changes the information set with all-node targets.

Method	Cora	PubMed	Roman-empire	Amazon-ratings
Supervised MLP	0.638 [0.578, 0.686]	0.122 [0.118, 0.127]	0.583 [0.579, 0.588]	0.196 [0.188, 0.204]
Soft KD (strict)	0.519 [0.451, 0.579]	0.101 [0.098, 0.105]	0.557 [0.552, 0.561]	**0.144 [0.135, 0.151]**
GLNN objective (strict)	0.593 [0.531, 0.651]	0.109 [0.106, 0.113]	0.566 [0.561, 0.570]	0.168 [0.159, 0.175]
KRD adapter (strict)	0.539 [0.473, 0.595]	0.114 [0.110, 0.119]	0.773 [0.768, 0.778]	0.172 [0.163, 0.180]
TINED adapter (strict)	**0.408 [0.365, 0.449]**	**0.095 [0.092, 0.100]**	**0.531 [0.525, 0.537]**	0.157 [0.151, 0.163]
GLNN (all-node targets)	0.050 [0.047, 0.052]	0.094 [0.092, 0.097]	0.466 [0.463, 0.469]	0.139 [0.134, 0.144]

Table 20: Mean real-graph performance at radius $k = 4$, averaged over teachers, five splits, and three seeds. Accuracy is evaluated against ground truth.

Dataset	Hop-MLP			Local GCN		
	KL↓	Agree↑	Acc.↑	KL↓	Agree↑	Acc.↑
Cora	0.0692	0.9413	0.8690	0.0626	0.9434	0.8589
PubMed	0.0249	0.9664	0.8856	0.0194	0.9709	0.8805
Roman-empire	0.2865	0.8308	0.7541	0.1307	0.8920	0.7946
Amazon-ratings	0.0454	0.8248	0.4568	0.0275	0.8590	0.4581

E Scope of the empirical audits

The coordinate-conditioned post-training audit establishes KL accounting under the declared fitted laws, not an intrinsic graph property or uncertainty across independent graphs. The held-out law diagnostics assess the tested families, not their truth or uniqueness. The expanded training-span calculation does not reconstruct the full optimization transcript. Width controls isolate a removable student component under matched summaries but do not establish student optimality. The real-topology and nonlinear audits evaluate distinct estimands; neither supplies a distribution-free local stopping rule.

F Data and Code Availability

Data availability. All datasets are public and are retrieved through PyTorch Geometric from the Planetoid and heterophilous-graphs sources cited in the paper. The five official split indices used in the experiments are specified in the protocol and included with the reproducibility metadata.

Code availability. The supplementary material includes the experiment, aggregation, plotting, and validation code; raw and aggregated results; run metadata; dependency versions; and the LaTeX source.