HoTS: Homophily-Aware Temperature Scaling for Graph Neural Network Calibration

Inwoo Tae
UNIST
vitainu0104@unist.ac.kr

Yoontae Hwang*
Pusan National University
yoontae.hwang@pusan.ac.kr

Yongjae Lee*
UNIST
LinqAlpha
yongjaelee@unist.ac.kr

Abstract

For graph node classification, calibrated class probabilities are needed when confidence scores, usually the maximum predicted class probability, are used to rank predictions, defer uncertain nodes to human review, or control risk. Existing post-hoc calibrators either apply one global temperature or use graph-aware modules without a principled structural form. We study how local graph structure should enter node-level calibration. Our first results show that a logit-only temperature rule is insufficient when nodes with identical logits but different local homophily require different optimal temperatures. We then analyze a population-concentration contextual stochastic block model with Gaussian features and a one-layer linear GCN. Under equidistant class means, the Bayes posterior over class-template scores is a temperature-scaled softmax whose inverse-temperature is governed by a homophily-dependent signal strength. In the positive-signal homophilic regime, the resulting temperature decreases approximately inversely with normalized local homophily. This law motivates Homophily-aware Temperature Scaling (HoTS), a simple post-hoc calibrator that assigns each node a positive scalar temperature from entropy-based logit concentration and estimated local homophily. HoTS has three temperature parameters, preserves the predicted class, and learns the strength of the structural correction from calibration data. Across 18 node-classification benchmarks, two GNN backbones, and eight calibration baselines, HoTS achieves the best mean Expected Calibration Error (ECE) of 4.79%, the best average rank, and the most reliable confidence ranking in selective classification. Code is available at https://github.com/inu0104/HoTS.

1 Introduction

Graph neural networks (GNNs) are widely used for node classification on relational data [Kipf and Welling, 2017, Hamilton et al., 2017]. In graph learning systems, probabilities often guide ranking, triage, human review, and risk scoring, not only class labels. Calibration, the agreement between predictive confidence and empirical correctness, is therefore central to reliability [Naeini et al., 2015]. Post-hoc calibration improves these probabilities without retraining the base model [Niculescu-Mizil and Caruana, 2005]. Temperature scaling is the standard approach, using a single positive scalar to rescale logits while preserving accuracy [Guo et al., 2017], but it assumes that miscalibration is sufficiently uniform across examples. This assumption is fragile on graphs. Through message passing, a node prediction depends on both its own features and the realized composition of its neighborhood. Thus, two nodes with the same logits or confidence can have different probabilities of being correct. A homophilic neighborhood may support a confident prediction, whereas a mixed or heterophilic neighborhood may make the same confidence unreliable [Zhu et al., 2020]. Since these regimes

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

can coexist in one graph, a global temperature must compromise across structurally different nodes. Recent graph-aware calibration methods address this issue with node-wise temperatures or graph-dependent logit transformations using auxiliary GNNs, attention, spectral features, or ensemble-style architectures [Wang et al., 2021, Hsu et al., 2022, Zhuang et al., 2025, Xie et al., 2024]. They show that graph structure matters, but do not specify the form of the structural correction.

This work asks *how the optimal node-wise temperature should depend on local graph structure, given that neighborhood composition affects the reliability of GNN confidence*. We show that logits alone are insufficient. Two nodes with the same logits can require different optimal temperatures when their local homophily provides different class-consistent support. Under this condition, homophily must enter the calibrator to avoid a residual calibration gap. We analyze a population-concentration Contextual Stochastic Block Model (CSBM) with Gaussian node features and a one-layer linear GCN with self-loops. CSBM permits an explicit Bayes posterior calculation while preserving the effect of neighborhood composition on the aggregated signal. Under equidistant class means, the Bayes posterior for class-template scores depends on local homophily through a homophily-dependent signal strength. In the positive-signal regime, stronger homophilic support sharpens the posterior, while weaker support softens it. In the strongly homophilic regime, the Bayes temperature decreases approximately in inverse proportion to normalized local homophily, so well-supported nodes need less correction than structurally ambiguous ones. Because the Bayes result uses oracle homophily, we convert it into a practical rule using estimated local homophily and entropy-based logit concentration. An errors-in-variables argument shows that noisy homophily estimates attenuate the inverse-homophily relationship, motivating us to learn the homophily sensitivity.

These results motivate Homophily-aware Temperature Scaling (HoTS), a node-wise calibrator. HoTS uses a global offset, a correction scale, and a homophily-sensitivity exponent to assign each node a positive scalar temperature from entropy-based logit concentration and estimated local homophily, preserving the predicted class whenever the top logit is unique. Across 18 datasets and eight baselines, HoTS achieves the lowest mean ECE of 4.79% and the most reliable confidence ranking in selective classification. Our contributions are to establish the necessity of homophily for optimal graph calibration, derive a CSBM Bayes calibration law linking temperature to local homophily, translate this law into a simple post-hoc method, and validate it across diverse benchmarks.

2 Related Work

Post-hoc calibration Post-hoc calibration maps model scores to reliable probabilities using a held-out calibration set while keeping the trained predictor fixed [Niculescu-Mizil and Caruana, 2005, Naeini et al., 2015]. Classical methods include Platt scaling for binary classification [Platt, 1999] and multiclass extensions via score decomposition or per-class probability estimation [Zadrozny and Elkan, 2002]. For deep neural networks, temperature scaling has become a standard baseline because it learns only one scalar temperature and preserves the predicted label whenever the top logit is unique [Guo et al., 2017]. Its simplicity makes it robust and easy to deploy, but limits its ability to correct instance-dependent calibration errors. More expressive calibrators address this limitation by learning richer transformations of logits or probabilities. Vector and matrix scaling introduce class-dependent or affine logit transformations [Guo et al., 2017], while beta calibration and Dirichlet calibration capture asymmetric and class-specific distortions [Kull et al., 2017, 2019]. Other approaches use ensembles, confidence-dependent temperatures, or entropy-based temperature functions [Zhang et al., 2020, Balanya et al., 2024]. These methods are effective in standard classification settings, but they usually treat examples as independent calibration instances and do not model how relational structure affects confidence reliability. A related issue is whether calibration should preserve the original prediction. General logit transformations can change the predicted class and mix calibration with refinement, whereas order-preserving calibrators avoid this [Rahimi et al., 2020]. Label-wise reliability reranking also addresses residual correctness differences after calibration while preserving predicted labels and class probabilities [Tae and Lee, 2026]. HoTS follows this principle by assigning each node a positive scalar temperature, preserving the predicted class while allowing the correction strength to vary with local graph structure.

Calibration for graph neural networks Calibration for graph neural networks is more challenging because node predictions are coupled through message passing. Empirical studies show that GNN calibration varies substantially across datasets and nodes, and that standard post-hoc methods often

fail to account for graph-specific reliability patterns [Teixeira et al., 2019]. Several works therefore incorporate graph structure into calibration. CaGCN [Wang et al., 2021] learns a topology-aware calibration function with an auxiliary GCN and exploits the observation that confidence patterns can be homophilic. RBS [Liu et al., 2022] groups nodes using a same-class-neighbor ratio and learns bin-wise temperatures. GATS [Hsu et al., 2022] analyzes sources of GNN miscalibration, including nodewise predictive diversity, distance to training nodes, relative confidence, and neighborhood similarity, and uses an attention-based model for node-wise temperature scaling. Recent work further expands the structural signals used for graph calibration. SimCalib [Tang et al., 2024] uses global and local nodewise similarity signals, ASTS [Xie et al., 2024] studies heterophily-aware calibration through spectral information, WATS [Li et al., 2025] uses graph wavelet features for node-specific temperature scaling, and GETS [Zhuang et al., 2025] combines input and model ensembles through a graph mixture-of-experts framework. Other methods improve calibration at training time or by modifying the graph itself. GCL [Wang et al., 2022] introduces a graph calibration loss, Moderate Message Passing [Wang et al., 2024] mitigates confidence bias induced by message passing, DCGC [Yang et al., 2024] improves calibration through data-centric edge weighting, and Fang et al. [2024] emphasize that calibration should preserve discriminative ability between correct and incorrect predictions. These studies establish that graph structure is crucial for calibration. However, most existing methods introduce structure through learned architectures, handcrafted features, spectral filters, wavelets, ensembles, or training objectives. They do not derive what functional relationship should hold between local graph structure and the optimal node-wise temperature. Our work fills this gap by deriving a homophily-dependent temperature law and translating it into a simple calibrator.

Graph structure theory and CSBM analysis The role of homophily and heterophily in GNNs has been widely studied for node classification. Prior work shows that standard message passing can struggle under heterophily and motivates architectures that better separate ego and neighbor information or exploit higher-order neighborhoods [Zhu et al., 2020]. Other studies show that the relationship between homophily and GNN performance is more nuanced than a global homophily measure suggests [Ma et al., 2021]. Recent work further emphasizes local homophily and shows that mismatch between local and global structure can create performance differences across node groups [Loveland et al., 2024]. These findings motivate our focus on node-level structural reliability. The contextual stochastic block model provides a principled framework for analyzing how graph structure and node features jointly affect GNN behavior. Foundational CSBM work characterizes thresholds for recovery in attributed graphs [Lu and Sen, 2023]. Recent analyses study Bayes-optimal inference and the performance of graph convolutional architectures under CSBM assumptions [Duranthon and Zdeborová, 2023, Dalle and Thiran, 2024]. These works clarify how structure, features, and message passing affect classification accuracy and representation quality. Our work is complementary to this theoretical literature. Existing CSBM analyses mainly study detection, classification, separability, or representation behavior, whereas calibration asks whether predicted probabilities match true correctness probabilities. To our knowledge, no prior work derives the Bayes-optimal temperature as an explicit function of local homophily under a CSBM. HoTS connects this CSBM Bayes posterior calculation to a practical node-wise temperature scaling rule.

3 Derivation of Node-wise Temperature

Assumptions. The theory is developed under a uniform-prior CSBM with Gaussian node features and a one-layer linear GCN with self-loops. For the exact posterior calculation, we use a population-concentration version of the model: conditional on the realized oracle local homophily h_i, the degree and neighbor class composition are fixed at their population values, so finite degree and composition fluctuations are not included in the posterior calculation. The class means are assumed to be equidistant, so that after centering they form a regular simplex on their affine span. Throughout the theoretical sections, h_i denotes the oracle local homophily defined using true labels. It is used only to characterize the Bayes signal and is not assumed to be available at test time. The implemented method uses an estimated homophily $\hat{h}_i$ from an auxiliary GCN supervised using observed labels. The formal model and posterior calculation are in Appendix A.1 and Appendix A.3.

Theoretical Motivation The theory has three roles. **First**, it shows that homophily is not merely a useful heuristic: if the Bayes-optimal temperature varies with h after conditioning on logits, then excluding h causes a positive calibration gap. **Second**, under the population-concentration CSBM, it

computes the Bayes posterior whose inverse-temperature is controlled by a homophily-dependent signal strength. **Third**, it turns this group-level Bayes signal into the HoTS form by adding a node-wise entropy proxy and a learnable exponent that accounts for noisy homophily estimation.

Why Homophily Is Necessary for GNN Calibration Let $\tau = 1/T$ denote inverse-temperature. For fixed logits z and label y, define the temperature-scaled cross-entropy loss as $\ell(\tau; z, y) = -\log \mathrm{softmax}(\tau z)_y$. Let $\mathcal{T}_z$ be the class of positive inverse-temperature calibrators depending only on z, and let $\mathcal{T}_{z,h}$ be the corresponding class of calibrators that may depend on both z and h.

Proposition 1 (Necessity of homophily)**.** *Let $\tau^*(z, h)$ be the CE-optimal positive inverse-temperature conditional on $(Z = z, H = h)$, and let $\tau^\star(z)$ be the CE-optimal positive inverse-temperature conditional only on $Z = z$. Under the local regularity conditions stated in Appendix A.2, if*

$$\mathrm{Var}_H(\tau^*(z, H) \mid Z = z) > 0 \tag{1}$$

on a set of positive measure, then

$$\inf_{\tau \in \mathcal{T}_{z,h}} \mathbb{E}[\ell(\tau(Z, H); Z, Y)] < \inf_{\tau \in \mathcal{T}_z} \mathbb{E}[\ell(\tau(Z); Z, Y)]. \tag{2}$$

Moreover, in the local quadratic regime, this improvement is proportional to the curvature-weighted conditional variance of $\tau^*(Z, H)$ given Z.

Proof. The proof is given in Appendix A.2. □

Intuitively, if two nodes have the same logits but different homophily values, and the Bayes-optimal temperature differs between them, then no logit-only calibrator can be optimal for both. Homophily is therefore necessary whenever it explains residual variation in the optimal positive temperature after conditioning on logits. Empirically, calibrators incorporating estimated local homophily (such as HoTS) outperform logit-only calibrators (TS, ETS) in aggregate across the 18 benchmarks (Section 5), which provides direct evidence for this gap. Proposition 1 reduces the question to whether the Bayes-optimal positive temperature genuinely varies with h within fixed-logit groups. The next result verifies the homophily dependence of the Bayes inverse-temperature for $Z = S$, where S denotes the CSBM class-template score vector, on the positive-signal region; it does not claim the same exact posterior for arbitrary learned logits. It computes the corresponding signed Bayes inverse-temperature, and ordinary positive-temperature statements are restricted to the positive-signal region $\mathcal{H}_+$. Thus Proposition 1 applies here with the logit variable specialized to the class-template score vector $Z = S$, provided that H remains non-degenerate within fixed-S groups, as formalized in the strictness argument in Appendix A.3.

CSBM Bayes Posterior and Homophily-Dependent Temperature Let $m_c = \mu_c - \bar{\mu}$ and $\rho_0 = \|m_c\|$. Under the equidistant-means assumption stated in Appendix A.1, the centered class means form a regular simplex on their affine span. To avoid confusion with the learned GCN logits, we denote the theoretical centered class-template scores by $s_c = \rho_0^{-1} m_c^\top (\bar{x}_i - \bar{\mu})$. The exact posterior result below is a statement about these class-template scores, not about an arbitrary learned linear head W; learned logits are connected to the theory only through the later surrogate HoTS form. We write the normalized homophily as $\tilde{h} = \frac{Kh-1}{K-1}$, so that $\tilde{h} = 0$ corresponds to a class-random neighborhood and $\tilde{h} > 0$ corresponds to homophily above the class-random baseline. Finally, let $v = \sigma^2/(\bar{d} + 1)$ and $\Gamma(h) = \rho_0(1 + \bar{d}\tilde{h})/(\bar{d} + 1)$.

Theorem 1 (CSBM Bayes posterior for class-template scores)**.** *Under the population-concentration CSBM described above and formalized in Appendix A.1, the Bayes posterior conditional on the realized homophily $H_i = h$ and the class-template score vector s is*

$$P(Y_i = c \mid s, H_i = h) = \mathrm{softmax}\left(\frac{\Gamma(h)}{v} s\right)_c. \tag{3}$$

Hence, for these class-template scores, the signed Bayes inverse-temperature is $\tau_{\mathrm{Bayes}}(h) = \Gamma(h)/v$.

Proof. The proof is given in Appendix A.3. □

Theorem 1 says that, under the population-concentration CSBM and for the class-template logits, local homophily changes the Bayes posterior only through the scalar signal strength $\Gamma(h)$. Consequently, for non-constant class-template scores s and $h \in \mathcal{H}_+$, the CE-optimal positive inverse-temperature conditional on (s, h) is $\Gamma(h)/v$. When $\Gamma(h) > 0$, this signed Bayes inverse-temperature corresponds to an ordinary positive temperature. When $\Gamma(h) < 0$, the Bayes rule reverses the template-logit ordering and cannot be represented by positive temperature scaling. Since $\Gamma(h)$ is increasing in h, more homophilic neighborhoods induce a larger Bayes inverse-temperature, and therefore a sharper posterior, whenever the signal is positive. On the positive-signal region $\mathcal{H}_+ = \{h : \Gamma(h) > 0\}$, the corresponding positive Bayes temperature is

$$T_{\mathrm{Bayes}}(h) = \frac{v}{\Gamma(h)} = \frac{\sigma^2}{(1 + \bar{d}\tilde{h})\rho_0}. \tag{4}$$

In the large-$\bar{d}\tilde{h}$ homophilic regime, $\Gamma(h) \approx \Gamma_0 \tilde{h}$, where $\Gamma_0 = \rho_0 \bar{d}/(\bar{d} + 1)$. Thus $T_{\mathrm{Bayes}}(h) \propto \tilde{h}^{-1}$, which motivates the homophily-dependent denominator of the HoTS temperature defined in Section 4.

From Bayes Temperature to a Node-wise Surrogate The Bayes temperature in (4) is group-level: all nodes with the same homophily receive the same temperature. To translate this group-level law into a node-wise rule, we use a reliability-matching surrogate. Consider binary classification, where node i has logits $z_i = (z_{i1}, z_{i2})$. When the predicted class is 1, $z_{i1} > z_{i2}$, so the positive logit gap $\delta_i = z_{i1} - z_{i2}$ is positive and $\sigma(\delta_i) \in (1/2, 1)$ is the model's confidence in the predicted class. Temperature scaling preserves this range: for any $T_i > 0$, $\sigma(\delta_i/T_i) \in (1/2, 1)$ as well. Calibration asks that this confidence equal the true correctness probability for node i. Since the per-node correctness probability is unobservable, we replace it with a group-level rate. Theorem 1 shows that the Bayes temperature depends on local homophily, so we group nodes by h and use the homophily-conditional correctness target $q(h_i) = \Pr(\hat{y}_i = y_i \mid H_i = h_i)$. Matching the calibrated confidence to this target, $\sigma(\delta_i/T_i) = q(h_i)$, and applying $\mathrm{logit} = \sigma^{-1}$ to both sides gives

$$T_i^{\mathrm{match}} = \frac{\delta_i}{\mathrm{logit}(q(h_i))}. \tag{5}$$

This is not an exact pointwise CE optimum under the full CSBM posterior; it is a practical rule that separates node confidence (through δ_i) from structural reliability (through $q(h_i)$). Equation (5) is well-defined and yields a positive temperature only when $q(h_i) > 1/2$; this corresponds to the positive-signal regime $h \in \mathcal{H}_+$ of Theorem 1, in which the Bayes posterior is sharper than chance. The implemented HoTS rule extends the form of this surrogate to all nodes via the absolute-value denominator $(|\hat{\tilde{h}}_i| + \epsilon)^\alpha$ (with $\hat{\tilde{h}}_i$ the normalized estimated homophily and $\epsilon > 0$ a small stabilizer) for stability when the estimated structural signal may be weak or negative, rather than as an exact pointwise approximation outside H_+.

For multi-class logits, we replace the binary gap by the entropy-based concentration proxy $\sqrt{2K \log K(1 - e_i)}$, where e_i is the normalized predictive entropy, obtained from the local softmax-entropy expansion around the uniform distribution. Under the binary centered class-template CSBM, the group correctness probability satisfies $q(h) = \Phi(\Gamma(h)/\sqrt{v})$. In the moderate-margin and large-$\bar{d}\tilde{h}$ homophilic regime, this gives $\mathrm{logit}(q(h)) \approx \lambda \tilde{h}$ for a positive constant λ. The entropy expansion and the binary reliability calculation are given in Appendix A.5 and Appendix A.6.

Combining these approximations yields, for $\tilde{h}_i > 0$ with $\bar{d}\tilde{h}_i \gg 1$, the positive-temperature oracle ratio form

$$T_i^{\mathrm{match}} \approx c\, \frac{\sqrt{2K \log K(1 - e_i)}}{\tilde{h}_i}, \qquad c > 0, \tag{6}$$

valid in the moderate-margin, large-$\bar{d}\tilde{h}$ part of the positive-signal homophilic regime. This oracle ratio is not intended for $\tilde{h}_i \le 0$; the absolute-value denominator in the implemented HoTS rule is a stable positive-temperature parametrization rather than an exact signed Bayes posterior approximation. Thus, temperature increases with raw logit concentration and decreases with structural reliability in this approximation. The practical form replaces the oracle $\tilde{h}_i$ by the estimated $\hat{\tilde{h}}_i$ and introduces a learnable exponent α on the homophily denominator $(|\hat{\tilde{h}}_i| + \varepsilon)^\alpha$ to account for estimation noise and deviations from the ideal CSBM regime.

Why the Exponent Is Learnable The large-$\bar{d}\tilde{h}$ approximation to the CSBM Bayes temperature gives the inverse-homophily law $T \propto \tilde{h}^{-1}$, corresponding to the oracle exponent $\alpha = 1$ in that regime. In practice, however, the oracle $\tilde{h}$ is replaced by a noisy estimate $\hat{\tilde{h}}$. This weakens the observed log-log relationship between temperature and homophily, so the exponent should be learned rather than fixed.

Proposition 2 (Exponent attenuation)**.** *Consider the regime where the positive-temperature oracle is approximated by an inverse-magnitude law. Let $u = \log(|\tilde{h}| + \varepsilon)$ and let the observed log-homophily be $\hat{u} = u + \eta$, where η is zero-mean noise independent of u with variance τ^2. If the oracle relation is $\log T^* = a - u$, but we fit the surrogate $\log T_\alpha = a_\alpha - \alpha\hat{u}$ by population least squares, then*

$$\alpha^* = \frac{\mathrm{Var}(u)}{\mathrm{Var}(u) + \tau^2} \leq 1. \tag{7}$$

The inequality is strict whenever $\tau^2 > 0$ and $\mathrm{Var}(u) > 0$.

Proof. The proof is given in Appendix A.7. □

Proposition 2 is exact for the log-space least-squares surrogate, not for the full CE objective used to train HoTS. On CSBM, the fitted α decreases from 1.009 ($\tau = 0$, oracle recovery) to 0.182 ($\tau = 0.5$), confirming the errors-in-variables attenuation (Figure 1b). Fixing $\alpha = 1$ on real data degrades mean ECE by $+1.10$ percentage points (Appendix B.9).

4 Homophily-Aware Temperature Scaling

Method We propose Homophily-aware Temperature Scaling (HoTS), a node-wise calibration method that combines predictive uncertainty with local graph structure. The theory suggests two design principles. First, in the large-$\bar{d}\tilde{h}$ part of the positive-signal homophilic CSBM regime, the Bayes temperature decreases approximately inversely with normalized homophily. Second, node-level confidence variation can be captured by an entropy-based logit-concentration proxy. For each node i, HoTS defines the temperature

$$T_i = T_{\mathrm{base}} + \frac{\beta\sqrt{2K\log K(1 - e_i)}}{(|\hat{\tilde{h}}_i| + \varepsilon)^\alpha} \tag{8}$$

where $e_i = H(\mathrm{softmax}(z_i))/\log K$ is the normalized predictive entropy, and $\hat{\tilde{h}}_i = (K\hat{h}_i - 1)/(K - 1)$ is the normalized estimated homophily. The calibrated prediction is $\mathrm{softmax}(z_i/T_i)$. Equation (8) acts on the learned GCN logits z_i, whereas Theorem 1 characterizes the Bayes posterior with respect to the CSBM class-template scores s; HoTS thus uses the learned logits as a practical surrogate for the unobserved class-template scores, and the form of the temperature is motivated by, rather than derived from, the exact Bayes rule. Section 5 reports the main calibration results across 18 benchmarks.

Homophily estimation The method uses the estimated homophily $\hat{h}_i$ rather than the oracle homophily h_i. In our implementation, $\hat{h}_i$ is produced by a small auxiliary GCN trained on targets computed using only training and validation labels, excluding self-loops and unobserved-label neighbors (architecture and training details in Appendix C.5). The predictor is held fixed during temperature fitting. This construction avoids test-label leakage while retaining local graph information.

Learnable parameters The learnable temperature parameters are $T_{\mathrm{base}} > 0$, $\beta > 0$, and $\alpha > 0$. The base temperature T_{base} provides a global positive offset, β controls the magnitude of the homophily-aware correction, and α controls the sensitivity to the estimated homophily magnitude. The stabilizer $\varepsilon > 0$ is fixed. The CSBM analysis (Proposition 2) suggests an asymptotic exponent in $[0, 1]$ under the log-space least-squares surrogate, but we do not enforce this upper bound in implementation, since real graphs may deviate from the idealized regime. In implementation, the positivity constraints on T_{base}, β, and α are enforced through softplus parameterizations.

Interpretation and prediction preservation The numerator $\sqrt{2K \log K(1-e_i)}$ increases with logit concentration and is small near the uniform predictive distribution. In the oracle positive-signal homophilic regime, a denominator $(\tilde{h}_i + \varepsilon)^\alpha$ is motivated by the large-$\tilde{dh}$ inverse-homophily Bayes temperature law, not by the exact finite-$\tilde{dh}$ Bayes expression near $\tilde{h}_i = 0$. In the implemented HoTS rule, we replace the oracle $\tilde{h}_i$ by the estimated signal and use $(|\hat{\tilde{h}}_i| + \varepsilon)^\alpha$ as a stable positive-temperature parametrization when the estimated structural signal may be negative, weak, or noisy. The offset T_{base} and stabilizer ε absorb the finite baseline behavior that is not captured by the large-signal inverse law. This absolute-value extension should not be interpreted as an exact signed Bayes posterior approximation in the negative-signal regime. In the idealized CSBM regime with oracle homophily, the asymptotic exponent is $\alpha = 1$. With estimated homophily and model mismatch, the exponent is learned from calibration data, allowing the strength of the homophily correction to adapt to each dataset. Since HoTS uses positive temperatures, it preserves the original predicted class and adjusts only the confidence assigned to the prediction.

5 Empirical Validation

The empirical study addresses three questions. We ask whether a homophily-aware temperature rule improves calibration across graph regimes without a high-capacity calibration network, whether controlled CSBM experiments recover the predicted inverse dependence of temperature on homophily and its attenuation under noisy homophily estimates, and whether HoTS yields confidence rankings that are more useful for downstream decisions.

Experimental Setup The evaluation covers 18 public node-classification benchmarks spanning citation graphs (Cora, CiteSeer, PubMed) [Yang et al., 2016], Amazon co-purchase and Coauthor graphs (Computers, Photo, CS, Physics, CoraFull) [Shchur et al., 2018], WebKB and Wikipedia-derived heterophilic graphs (Texas, Cornell, Wisconsin, Chameleon, Squirrel, Actor) [Pei et al., 2020], the heterophilous benchmark of Platonov et al. [2023] (Roman-Empire, tolokers), and large-scale graphs ogbn-arxiv [Hu et al., 2020] and Reddit [Hamilton et al., 2017]. Following convention in prior GNN and calibration benchmarks, Table 1 groups these datasets into homophilic, heterophilic, and large-scale regimes; the heterophilic group follows the benchmark designations of Pei et al. [2020] and Platonov et al. [2023] rather than a strict homophily threshold, and per-dataset homophily statistics are reported in Appendix C.1. Unless otherwise stated, each dataset and backbone pair is run with 10 random seeds under a $20/10/70$ train, validation, and test split. The backbones are GCN [Kipf and Welling, 2017] and GAT [Veličković et al., 2017]. Table 1 pools both backbones, while Appendix B.2 gives the per-backbone breakdown. The primary metric is Expected Calibration Error (ECE) with 15 equally spaced confidence bins. Because ECE is a binned summary, Appendix B.1 also reports negative log-likelihood (NLL) and degree-stratified ECE. Baselines include standard post-hoc calibrators, Temperature Scaling (TS), Vector Scaling (VS) [Guo et al., 2017], Ensemble Temperature Scaling (ETS) [Zhang et al., 2020], entropy-aware HTS [Balanya et al., 2024], and graph-aware calibrators CaGCN, GATS, GETS, and WATS. RBS and SimCalib are omitted only because their public reference could not be located, while Section 2 discusses them as the closest prior works to HoTS. GATS exceeds the 24 GB GPU memory budget on Reddit and is therefore reported as "–" there (Appendix C.2). HoTS fits its three temperature parameters on the calibration split, uses training-split cross-entropy for early stopping, and uses the homophily estimate described in Section 4.

Calibration Across Graph Regimes Table 1 reports test ECE. HoTS achieves the lowest mean ECE, 4.79%, and the best average rank, 3.33, among the compared calibrators. On several homophilic datasets, the strongest baselines are statistically close. On PubMed, CoraFull, Roman-Empire, and tolokers, another method obtains the lowest ECE. HoTS improves the aggregate calibration profile while using a much smaller calibrator, and provides substantial improvements on benchmarks such as Texas, Cornell, and Wisconsin. This is consistent with the motivation of HoTS. A single global temperature can be sufficient when reliability errors are nearly uniform, whereas graph-dependent correction becomes more useful when local structure changes the meaning of confidence. Results under NLL and degree-stratified ECE are reported in Appendix B.1.

Controlled Validation of the Structural Law The theory predicts two qualitative effects. Theorem 1 implies that, in the positive-signal CSBM regime, the Bayes temperature decreases as the

Table 1: Calibration ECE (%, mean±std over 10 seeds per backbone, pooled across GCN and GAT). Best per row in red, second in **bold**, third underlined (excluding Uncal).

Dataset	Uncal	TS	VS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS (ours)
Homophilic										
Cora	19.64 ± 2.62	3.14 ± 0.82	$\mathbf{2.90 \pm 0.78}$	2.94 ± 0.82	3.58 ± 0.56	4.12 ± 0.78	2.99 ± 0.96	3.86 ± 0.82	3.08 ± 0.73	2.50 ± 0.72
CiteSeer	18.57 ± 4.19	$\mathbf{3.04 \pm 0.62}$	3.20 ± 0.56	2.93 ± 0.63	3.78 ± 0.86	4.81 ± 0.79	3.19 ± 0.86	5.08 ± 1.07	3.10 ± 0.57	3.50 ± 0.80
PubMed	12.15 ± 2.15	1.04 ± 0.28	1.16 ± 0.39	1.00 ± 0.31	1.22 ± 0.65	1.11 ± 0.99	$\mathbf{0.88 \pm 0.28}$	1.14 ± 0.38	0.83 ± 0.30	1.26 ± 0.30
CoraFull	32.93 ± 4.51	2.76 ± 0.81	2.84 ± 0.63	2.20 ± 0.63	3.94 ± 0.63	$\mathbf{2.56 \pm 0.61}$	2.77 ± 0.80	3.72 ± 0.98	3.85 ± 1.65	2.73 ± 0.81
Computers	5.86 ± 1.42	2.19 ± 0.45	2.07 ± 0.45	2.12 ± 0.50	1.78 ± 0.55	1.39 ± 0.34	1.91 ± 0.49	2.66 ± 0.73	$\mathbf{1.62 \pm 0.60}$	1.76 ± 0.49
Photo	4.31 ± 0.90	1.23 ± 0.31	1.27 ± 0.26	1.30 ± 0.49	1.14 ± 0.48	1.35 ± 0.42	1.24 ± 0.32	1.42 ± 0.42	1.23 ± 0.39	$\mathbf{1.21 \pm 0.44}$
CS	2.72 ± 1.20	2.26 ± 0.83	2.03 ± 0.68	2.22 ± 0.87	1.34 ± 0.31	2.39 ± 0.83	2.11 ± 0.67	2.64 ± 2.82	3.25 ± 0.59	$\mathbf{1.51 \pm 0.60}$
Physics	1.20 ± 0.63	0.87 ± 0.29	0.85 ± 0.26	0.82 ± 0.31	0.52 ± 0.20	0.72 ± 0.26	0.72 ± 0.29	1.00 ± 0.65	1.41 ± 0.47	$\mathbf{0.54 \pm 0.23}$
Heterophilic										
Texas	21.86 ± 8.25	21.64 ± 8.49	18.23 ± 6.93	19.27 ± 8.41	$\mathbf{17.69 \pm 6.71}$	20.33 ± 7.82	17.85 ± 7.36	18.28 ± 4.98	24.78 ± 8.18	15.37 ± 5.90
Cornell	20.13 ± 8.13	20.05 ± 7.39	19.65 ± 6.58	18.18 ± 7.60	18.06 ± 7.39	$\mathbf{15.93 \pm 7.01}$	18.58 ± 7.91	19.58 ± 7.90	25.17 ± 6.87	14.28 ± 5.82
Wisconsin	19.31 ± 7.31	19.06 ± 7.19	19.62 ± 7.19	$\mathbf{16.81 \pm 6.81}$	18.13 ± 6.77	18.77 ± 6.76	17.83 ± 6.71	19.93 ± 7.08	24.57 ± 7.34	14.67 ± 5.38
Chameleon	11.50 ± 2.51	11.58 ± 2.57	11.24 ± 2.78	11.19 ± 3.63	8.95 ± 1.79	9.20 ± 2.38	10.67 ± 3.05	11.00 ± 3.43	13.18 ± 1.94	$\mathbf{9.06 \pm 2.35}$
Squirrel	6.19 ± 2.43	6.19 ± 2.42	6.17 ± 2.53	6.37 ± 2.61	$\mathbf{4.47 \pm 1.99}$	4.38 ± 1.60	5.87 ± 2.73	7.31 ± 3.04	6.55 ± 2.11	4.97 ± 1.53
Actor	2.47 ± 0.98	2.47 ± 0.84	1.96 ± 0.76	2.47 ± 0.93	2.68 ± 1.01	$\mathbf{2.37 \pm 0.87}$	2.52 ± 1.01	4.38 ± 2.13	4.77 ± 1.89	2.49 ± 0.85
Roman-Empire	12.27 ± 1.09	$\mathbf{2.64 \pm 0.65}$	2.10 ± 0.83	2.72 ± 0.77	4.06 ± 1.20	2.80 ± 0.84	2.84 ± 0.64	3.37 ± 0.95	2.93 ± 1.04	4.36 ± 0.78
tolokers	4.30 ± 1.09	4.22 ± 0.84	3.66 ± 0.81	4.23 ± 0.81	2.92 ± 0.95	$\mathbf{2.67 \pm 0.81}$	3.75 ± 0.93	2.04 ± 0.63	4.25 ± 0.65	3.25 ± 0.78
Large										
ogbn-arxiv	5.06 ± 1.99	2.65 ± 0.52	2.72 ± 0.34	2.57 ± 0.55	1.29 ± 0.18	1.82 ± 0.36	2.63 ± 0.77	4.51 ± 0.85	2.19 ± 0.58	$\mathbf{1.53 \pm 0.40}$
Reddit	5.35 ± 1.59	1.68 ± 0.67	1.84 ± 0.58	1.73 ± 0.52	$\mathbf{1.14 \pm 0.13}$	0.94 ± 0.48	–	1.92 ± 1.31	1.24 ± 0.47	1.25 ± 0.18
Avg	11.43	6.04	5.75	5.62	**5.37**	5.43	5.79	6.32	7.11	4.79
Avg rank	–	5.61	5.17	4.58	**3.94**	4.08	4.38	7.22	6.42	3.33

homophily signal $\Gamma(h)$ increases. Proposition 2 predicts that noisy homophily estimates attenuate the apparent power-law exponent. Figure 1 tests these two predictions in controlled CSBM experiments.

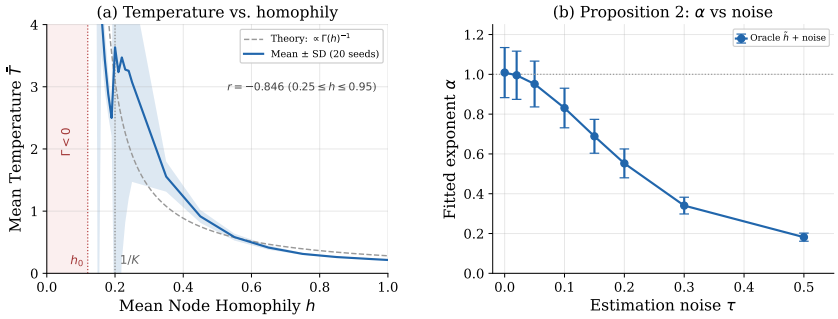

Figure 1: Controlled CSBM validation. **(a)** Fitted temperature over the full homophily range; the blue curve and shaded band show the mean and one standard deviation across 20 seeds. The dashed curve shows the inverse-homophily trend. Pearson correlation is computed over the eight sweep points in $0.25 \leq h \leq 0.95$. We mark h_0 and $1/K$ and shade $\Gamma(h) < 0$; the displayed vertical range is truncated. **(b)** The fitted exponent α is near the oracle value under noiseless homophily and decreases as estimation noise increases, matching Proposition 2's attenuation.

In the homophily sweep, all other CSBM parameters are fixed. Over the sweep range $0.25 \leq h \leq 0.95$, the mean fitted temperature decreases with homophily, with Pearson correlation $r = -0.846$ computed from the eight mean temperatures over 20 seeds, matching the predicted trend $T^*(h) \propto \Gamma(h)^{-1}$. Figure 1(a) also shows the additional experiments at lower homophily. In the noise sweep, we add controlled noise to the oracle homophily signal before fitting the power-law exponent. The fitted exponent decreases from 1.009 at zero noise to 0.182 at the largest noise level, consistent with the errors-in-variables attenuation formula. These experiments do not prove that real graphs follow the CSBM exactly. Instead, they verify that the structural dependencies used by HoTS are recoverable in the setting where the theory is meant to apply. On real benchmarks, fixing the exponent to the oracle value $\alpha = 1$ increases mean ECE by $+1.10$ percentage points under the matched comparison in Appendix B.9, supporting the decision to learn α from calibration data.

Selective Classification Calibration also guides when to trust a prediction. We therefore evaluate selective classification. Nodes are ranked by calibrated confidence, and the least confident nodes are rejected. Because HoTS uses positive scalar temperatures, it preserves the predicted class at full coverage. Any improvement below 100% coverage comes from a better confidence ranking rather

Table 2: Selective classification (top prediction-preserving baselines): retained-node accuracy (%) at coverage c, macro-averaged over 18 datasets, 2 backbones, and 10 seeds. Restricted to baselines that strictly preserve the predicted class (see Appendix C.6). Best per row in **bold** (excluding Uncal).

Coverage	Uncal	TS	HTS	HoTS (ours)
100%	67.39	67.39	67.39	67.39
95%	68.87	68.90	68.92	**68.94**
90%	70.01	70.08	70.13	**70.15**
85%	71.04	71.12	71.17	**71.18**
80%	71.94	72.05	**72.13**	72.11
75%	72.81	72.92	72.94	**72.96**
70%	73.60	73.73	73.75	**73.81**

than from changing the classifier. HoTS gives the highest retained accuracy among the compared methods at most evaluated coverage levels (Table 2). These results support the use of homophily-aware temperature scaling when downstream decisions depend on separating reliable predictions from less reliable ones.

Does Homophily Explain Residual Reliability? Proposition 1 motivates examining whether local homophily is associated with reliability beyond the base-model confidence. We stratify test nodes by confidence and compare empirical accuracy across oracle-homophily groups within each bin. Using oracle homophily lets us examine the reliability information in neighborhood label composition itself, separately from errors in estimating it. The high-homophily group is more accurate on most datasets, with pronounced gaps on homophilic and large-scale graphs and mixed behavior on heterophilic graphs (Figure 8). This retrospective diagnostic motivates homophily as a structural variable for calibration. True labels are used only to form the diagnostic groups; HoTS uses estimated homophily in its temperature map.

Additional Experiments Additional experiments are reported in the appendix to keep the main text focused. Additional GNN backbone comparisons are documented in Appendix B.3. Appendix B.1 and Appendix B.2 show robustness under NLL, degree-stratified ECE, and separate GCN and GAT evaluations. Appendix B.4 shows that HoTS reduces the residual calibration gap after temperature scaling and relates node-level confidence gaps to entropy concentration and oracle homophily. Appendix B.5 validates the sparse-overlap KL scaling used in the theory, and Appendix B.6 studies random edge perturbations. Appendix B.7 extends the residual-reliability diagnostic of this section. Appendix B.8 compares homophily with representative motif, centrality, and curvature statistics after controlling for confidence and degree. Appendix B.9 compares HoTS with entropy-only scaling to assess the contribution of homophily.

6 Conclusion

This paper presents a structural view of post-hoc calibration for graph neural networks. The main insight is that confidence cannot always be calibrated from logits alone, because the reliability of a graph prediction also depends on the local evidence induced by message passing. When nodes with comparable logits differ in neighborhood homophily, they can require different temperatures to produce reliable probabilities. We formalize this through a necessity result showing that a logit-only calibrator incurs an avoidable calibration gap whenever homophily explains residual variation in the optimal temperature. The CSBM analysis makes this principle explicit by showing that, for class-template scores under a population-concentration model, the Bayes inverse-temperature is governed by a homophily-dependent signal strength. In the positive-signal homophilic regime, this yields an approximate inverse relation between temperature and normalized local homophily.

HoTS translates this law into a three-parameter temperature map by combining entropy-based logit concentration, estimated homophily, and a learned exponent that adapts to noisy structural estimates. Since HoTS uses positive scalar temperatures, it improves probability reliability while preserving the predicted class. Across 18 benchmarks and both GCN and GAT backbones, HoTS achieves the best mean ECE and average rank among the compared calibrators. However, the exact Bayes characterization is derived under an idealized CSBM with Gaussian features, equidistant class means,

uniform class priors, and a one-layer linear GCN in a population-concentration limit. Thus, the inverse-homophily law should be viewed as a structural prior that motivates HoTS rather than an exact Bayes rule for arbitrary real graphs, especially under class imbalance, nonlinear multi-layer message passing, or low-degree finite-sample fluctuations. Overall, the results show that a compact theory-guided structural correction can provide an interpretable and effective path toward reliable probabilistic graph learning.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST)); the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528 and RS-2025-24803208); and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2026-25576625).

References

Sergio A Balanya, Juan Maroñas, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14):8073–8095, 2024.

Guillaume Dalle and Patrick Thiran. Optimal performance of graph convolutional networks on the contextual stochastic block model. In *The Third Learning on Graphs Conference*, 2024.

Odilon Duranthon and Lenka Zdeborová. Optimal inference in contextual stochastic block models. *arXiv preprint arXiv:2306.07948*, 2023.

Yujie Fang, Xin Li, Qianyu Chen, and Mingzhong Wang. Improving gnn calibration with discriminative ability: insights and strategies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11953–11960, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? *Advances in Neural Information Processing Systems*, 35:13775–13786, 2022.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

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

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Xiaoyang Li, Linwei Tao, Haohui Lu, Minjing Dong, Junbin Gao, and Chang Xu. Calibrating graph neural networks with wavelet-aware temperature scaling. *arXiv preprint arXiv:2506.23782*, 2025.

Tong Liu, Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Hang Li, and Volker Tresp. On calibration of graph neural networks for node classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

Donald Loveland, Jiong Zhu, Mark Heimann, Benjamin Fish, Michael T Schaub, and Danai Koutra. On performance discrepancies across local homophily levels in graph neural networks. In *Learning on Graphs Conference*, pages 6–1. PMLR, 2024.

Chen Lu and Subhabrata Sen. Contextual stochastic block model: Sharp thresholds and contiguity. *Journal of Machine Learning Research*, 24(54):1–34, 2023.

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

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Hongbin Pei, Bingzhe Wei, Kevin Chen Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *8th International Conference on Learning Representations*, 2020.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in neural information processing systems*, 33:13456–13467, 2020.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Inwoo Tae and Yongjae Lee. Post-calibration reliability reranking of relevance decisions via label-wise monotone projection. *arXiv preprint arXiv:2608.10406*, 2026.

Boshi Tang, Zhiyong Wu, Xixin Wu, Qiaochu Huang, Jun Chen, Shun Lei, and Helen Meng. Simcalib: Graph neural network calibration based on similarity between nodes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15267–15275, 2024.

Leonardo Teixeira, Brian Jalaian, and Bruno Ribeiro. Are graph neural networks miscalibrated? *arXiv preprint arXiv:1905.02296*, 2019.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Min Wang, Hao Yang, and Qing Cheng. Gcl: Graph calibration loss for trustworthy graph neural network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 988–996, 2022.

Min Wang, Hao Yang, Jincai Huang, and Qing Cheng. Moderate message passing improves calibration: A universal way to mitigate confidence bias in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21681–21689, 2024.

Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34:23768–23779, 2021.

Xiaotian Xie, Biao Luo, Yingjie Li, Chunhua Yang, and Weihua Gui. Exploring heterophily in calibration of graph neural networks. *Neurocomputing*, 604:128294, 2024.

Cheng Yang, Chengdong Yang, Chuan Shi, Yawen Li, Zhiqiang Zhang, and Jun Zhou. Calibrating graph neural networks from a data-centric perspective. In *Proceedings of the ACM Web Conference 2024*, pages 745–755, 2024.

Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR, 2016.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pages 11117–11128, 2020.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

Dingyi Zhuang, Chonghe Jiang, Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. GETS: Ensemble temperature scaling for calibration in graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Supplementary Material

Appendix Table of Contents

Appendix A	**Auxiliary Results and Proofs**
	A.1 Formal Population Concentration Model for Theorem 1
	A.2 Proof of Proposition 1
	A.3 Proof of Theorem 1
	A.4 Class-Symmetric Representative and Scalar h Sufficiency
	A.5 Entropy and Logit Taylor Expansions
	A.6 Binary CSBM Reliability Calculation
	A.7 Proof of Proposition 2
	A.8 Sparse-Overlap Gaussian Residual Approximation
Appendix B	**Additional Empirical Results**
	B.1 Additional Calibration Metrics
	B.2 Per-Backbone Results
	B.3 Calibration Across GNN Backbones
	B.4 Reliability Diagrams and Node-Level Diagnostics
	B.5 KL Bound Validation
	B.6 Structural Perturbation Stress Test
	B.7 Confidence-Conditioned Homophily Diagnostics
	B.8 Comparison with Other Structural Statistics
	B.9 HoTS Component Ablation
Appendix C	**Implementation Details**
	C.1 Dataset Statistics and Preprocessing
	C.2 Hardware and Software
	C.3 GNN Training Hyperparameters
	C.4 Calibration Method Hyperparameters
	C.5 HoTS Homophily Predictor
	C.6 Prediction Preservation Across Calibrators

A Auxiliary Results and Proofs

This appendix collects the technical details omitted from the main text. The results are organized according to the order in which they are used. We first state the population-concentration CSBM used for the posterior calculation. We then prove the homophily necessity result, the CSBM posterior theorem, and the class-symmetric scalar-sufficiency statement. We next give the entropy expansion, the binary reliability calculation, and the exponent attenuation proof. The final theoretical subsection states the sparse-overlap approximation used to justify the neighbor-logit product approximation.

A.1 Formal Population Concentration Model for Theorem 1

The posterior calculation in Theorem 1 uses the following population-concentration version of the CSBM. The graph has K classes with uniform prior. Edges are generated independently with probabilities p_{in} within classes and p_{out} across classes, with $p_{\text{in}} \neq p_{\text{out}}$. Node features satisfy

$$x_j = \mu_{y_j} + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, \sigma^2 I) \tag{9}$$

independently, with $\sigma > 0$.

The one layer aggregate is

$$\bar{x}_i = \frac{1}{d_i + 1} \left(x_i + \sum_{j \in \mathcal{N}(i)} x_j \right). \tag{10}$$

For the posterior calculation, conditional on $Y_i = c$ and the realized oracle local homophily $H_i = h$, we replace the degree by its population value $d_i = \bar{d}$ and fix the neighbor class composition at its population-concentration value. A fraction h of neighbors belongs to class c, and the remaining fraction $1 - h$ is class balanced over the other $K - 1$ classes. This deterministic composition approximation removes finite class composition fluctuations. The only remaining randomness in $\bar{x}_i$ comes from the Gaussian feature noises.

The class means are equidistant. There exists $c_\mu > 0$ such that

$$\|\mu_k - \mu_\ell\|_2 = c_\mu, \qquad k \neq \ell. \tag{11}$$

Let

$$\bar{\mu} = \frac{1}{K} \sum_{r=1}^{K} \mu_r, \qquad m_c = \mu_c - \bar{\mu}, \qquad \rho_0 = \|m_c\|. \tag{12}$$

Then $\{m_c\}_{c=1}^K$ forms a regular simplex on its affine span. We define the class-template score vector by

$$s_c = \rho_0^{-1} m_c^\top (\bar{x}_i - \bar{\mu}). \tag{13}$$

Finally, write

$$\tilde{h} = \frac{Kh - 1}{K - 1}, \qquad v = \frac{\sigma^2}{\bar{d} + 1}, \qquad \Gamma(h) = \frac{\rho_0 (1 + \bar{d}\tilde{h})}{\bar{d} + 1}. \tag{14}$$

A.2 Proof of Proposition 1

For each (z, h), let

$$r_{z,h}(\tau) = \mathbb{E}[\ell(\tau; Z, Y) \mid Z = z, H = h]. \tag{15}$$

Assume that $r_{z,h}$ has a unique interior minimizer $\tau^*(z, h)$. Let

$$\kappa(z, h) = \partial_\tau^2 r_{z,h}(\tau)\big|_{\tau = \tau^*(z,h)}. \tag{16}$$

Define the curvature-weighted projection

$$\bar{\tau}_\kappa(z) = \frac{\mathbb{E}_H[\kappa(z, H)\tau^*(z, H) \mid Z = z]}{\mathbb{E}_H[\kappa(z, H) \mid Z = z]}. \tag{17}$$

Assume that the within-z variation of $\tau^*(z, H)$ is locally small and that the third derivatives of $r_{z,h}$ are uniformly bounded in the relevant neighborhood. Let

$$\epsilon_z = \sup_h |\tau^*(z, h) - \bar{\tau}_\kappa(z)| \tag{18}$$

over the support of $H \mid Z = z$. Then

$$\begin{aligned}&\inf_{\tau \in \mathcal{T}_z} \mathbb{E}[\ell(\tau(Z); Z, Y)] - \inf_{\tau \in \mathcal{T}_{z,h}} \mathbb{E}[\ell(\tau(Z, H); Z, Y)] \\ &\quad = \frac{1}{2}\mathbb{E}\Big[\kappa(Z, H)\big(\tau^*(Z, H) - \bar{\tau}_\kappa(Z)\big)^2\Big] + o\big(\mathbb{E}_Z[\epsilon_Z^2]\big)\,.\end{aligned} \tag{19}$$

For fixed z, the loss is convex in τ because

$$\partial_\tau^2 \ell(\tau; z, y) = \mathrm{Var}_{p_\tau}(z_k) \geq 0, \tag{20}$$

where $p_\tau = \mathrm{softmax}(\tau z)$. It is strictly convex whenever z is not proportional to $\mathbf{1}$.

The displayed expansion follows by Taylor expanding $r_{z,h}$ around $\tau^*(z, h)$ and then minimizing the leading quadratic term over a single logit-only value. The minimizer of this quadratic approximation is precisely $\bar{\tau}_\kappa(z)$. Equivalently, $\bar{\tau}_\kappa(z)$ is the local-quadratic surrogate for the exact CE-optimal logit-only inverse-temperature $\tau^\star(z)$ in Proposition 1: under the bounded-third-derivative regularity above, $\tau^\star(z) = \bar{\tau}_\kappa(z) + o(\epsilon_z)$, so the displayed gap holds with $\tau^\star(z)$ in place of $\bar{\tau}_\kappa(z)$ at the same leading order. Hence the leading-order gap is a curvature-weighted conditional variance. If $\tau^*(z, H)$ is not almost surely constant given $Z = z$, strict convexity implies that no single logit-only value can attain all conditional minima simultaneously. This gives a strictly positive expected CE gap.

A.3 Proof of Theorem 1

Condition on $Y_i = c$ and $H_i = h$. Under the population-concentration model in Appendix A.1, the self-loop contributes one copy of class c, the same class neighbors contribute $h\bar{d}$ population weighted copies of class c, and the off class neighbors are class-balanced over the other $K - 1$ classes. Hence the conditional aggregate mean is

$$\bar{\mu} + a(h)m_c, \qquad a(h) = \frac{1 + \bar{d}\widetilde{h}}{\bar{d} + 1}. \tag{21}$$

Since the class composition is fixed at its population value, the only remaining randomness is the independent Gaussian feature noise. The aggregate covariance is vI, with $v = \sigma^2/(\bar{d} + 1)$. Thus

$$\bar{x}_i \mid (Y_i = c, H_i = h) \sim \mathcal{N}(\bar{\mu} + a(h)m_c, vI). \tag{22}$$

Decompose $\bar{x}_i - \bar{\mu}$ into its projection onto $\mathrm{span}\{m_1, \ldots, m_K\}$ and its orthogonal component. The orthogonal component has a class independent Gaussian distribution. The likelihood ratio depends on $\bar{x}_i$ only through

$$\big(m_1^\top(\bar{x}_i - \bar{\mu}), \ldots, m_K^\top(\bar{x}_i - \bar{\mu})\big)\,, \tag{23}$$

or equivalently through the class-template score vector s. Hence the posterior conditional on the full aggregate is the same function of s as the posterior conditional on s alone. Since the class prior is uniform and the model is class-symmetric, $P(Y_i = c \mid H_i = h) = 1/K$. Bayes rule gives a posterior proportional to

$$\exp\left(-\frac{\|\bar{x}_i - \bar{\mu} - a(h)m_c\|^2}{2v}\right). \tag{24}$$

Expanding the squared norm, the terms $\|\bar{x}_i - \bar{\mu}\|^2$ and $a(h)^2\|m_c\|^2$ are class independent and cancel in the softmax. The remaining term is

$$\frac{a(h)}{v}(\bar{x}_i - \bar{\mu})^\top m_c. \tag{25}$$

Since $s_c = \rho_0^{-1} m_c^\top(\bar{x}_i - \bar{\mu})$ and $\Gamma(h) = a(h)\rho_0$, we obtain

$$P(Y_i = c \mid s, H_i = h) = \mathrm{softmax}\left(\frac{\Gamma(h)}{v}s\right)_c. \tag{26}$$

Finally, if z has a unique maximum coordinate $\hat{y}$, then $\mathrm{softmax}(\tau z)_{\hat{y}}$ is strictly increasing in τ, because its derivative is

$$p_{\hat{y}}\left(z_{\hat{y}} - \sum_k p_k z_k\right) > 0. \tag{27}$$

Since $\Gamma(h)$ is strictly increasing in h, the posterior correctness probability is strictly increasing in h on

$$\mathcal{H}_+ = \{h \mid \Gamma(h) > 0\}. \tag{28}$$

Strictness of the homophily gap Restricted to the positive-signal subpopulation $H \in \mathcal{H}_+$, the CSBM posterior verifies the strictness condition in Proposition 1. If the conditional distribution of H on this subpopulation has at least two positive probability values in $\mathcal{H}_+$, then the Gaussian class-template score densities induced by those homophily values have overlapping support on the centered score subspace. Hence $P(H = h \mid S = s, H \in \mathcal{H}_+)$ is non-degenerate for almost every non-degenerate s. Since $\tau_{\mathrm{Bayes}}(h) = \Gamma(h)/v$ is strictly increasing in h on $\mathcal{H}_+$, the conditional variance of $\tau_{\mathrm{Bayes}}(H)$ given $S = s$ is positive for almost every such s. Therefore, under the population-concentration CSBM posterior, the positive temperature calibration gap in Proposition 1 is strict whenever homophily varies within fixed class-template score groups on $\mathcal{H}_+$.

Assume that H has at least two positive probability values $h \neq h'$ in $\mathcal{H}_+$. For each such value, the conditional distribution of the centered aggregate is a non-degenerate Gaussian in feature space because $\sigma > 0$. After projection onto the centered class-template logit subspace, the induced conditional density of Z is strictly positive on that subspace. Therefore the densities $f_{Z|H=h}(z)$ and $f_{Z|H=h'}(z)$ overlap on a set of full measure in the centered logit subspace.

By Bayes rule, both $P(H = h \mid Z = z)$ and $P(H = h' \mid Z = z)$ are positive for almost every such z. Hence H is not a deterministic function of Z. Since $\tau_{\mathrm{Bayes}}(h) = \Gamma(h)/v$ is strictly increasing in h on $\mathcal{H}_+$, the conditional distribution of $\tau_{\mathrm{Bayes}}(H)$ given $Z = z$ is non-degenerate for almost every non-degenerate z. This proves the strictness claim used after Theorem 1.

For non-atomic or binned homophily distributions, the same proof applies after replacing the two atoms by two disjoint measurable subsets of $\mathcal{H}_+$ with positive probability and overlapping conditional logit densities.

Positive and negative signal At $\tilde{h} = 0$, corresponding to a random neighborhood, $\Gamma(h) = \rho_0/(\bar{d} + 1) > 0$. Thus the node's own feature contributes signal even when the neighborhood is uninformative. The approximation $\Gamma(h) \approx \Gamma_0 \tilde{h}$ is valid only in the large-degree homophilic regime $\bar{d}\tilde{h} \gg 1$. If $1 + \bar{d}\tilde{h} < 0$, then $\Gamma(h) < 0$ and the signed Bayes inverse-temperature is negative. A negative inverse-temperature reverses the template logit ordering and cannot be represented by ordinary positive temperature scaling. Therefore exact positive temperature Bayes optimality statements are restricted to $\mathcal{H}_+$. The implemented HoTS method uses an absolute value extension for stability outside this regime. This extension should be interpreted as a calibration parametrization rather than as a signed Bayes posterior approximation.

Scope of the scalar temperature Under the population-concentration CSBM with equidistant class means, the centered class-symmetric representative makes the scalar group level Bayes temperature depend on the graph only through h. This is a statement about the idealized group level scalar temperature. It does not imply that arbitrary graph-aware features, such as neighbor features, neighbor logits, or finite-degree composition fluctuations, are conditionally irrelevant in finite CSBM or in real graphs. The symmetry argument is provided in Appendix A.4, and the sparse-overlap neighbor logit approximation is given in Appendix A.8.

A.4 Class-Symmetric Representative and Scalar h Sufficiency

This subsection records the symmetry argument behind the statement that, in the population-concentration CSBM, the scalar group level temperature depends on the graph only through h.

Let $m_c = \mu_c - \bar{\mu}$. Under the equidistant means assumption in Appendix A.1, the set $\{m_c\}_{c=1}^K$ is a regular simplex on its span. For every class permutation $\pi \in S_K$, there is an orthogonal map R_π on the simplex span such that $R_\pi m_c = m_{\pi(c)}$. Let P_π be the corresponding permutation matrix on output coordinates. Because the theoretical head acts on centered features, the CSBM distribution is invariant under the simultaneous transformation

$$(y, x - \bar{\mu}) \mapsto (\pi(y), R_\pi(x - \bar{\mu})). \tag{29}$$

Hence the population CE risk is invariant under

$$W \mapsto P_\pi W R_\pi^{-1}. \tag{30}$$

The population CE risk is convex in the linear classifier W. Therefore, if W^* is a population minimizer, the symmetrized classifier

$$W^\dagger = |S_K|^{-1} \sum_{\pi \in S_K} P_\pi W^* R_\pi^{-1} \tag{31}$$

is also a minimizer.

The symmetrized classifier is class equivariant, meaning that

$$W^\dagger R_\pi = P_\pi W^\dagger. \tag{32}$$

On the centered simplex span, this implies a one versus rest symmetric form, up to a softmax irrelevant constant shift of all logits. Under degree and off class composition concentration, the conditional distribution of logits given $(H_i = h, Y_i = c)$ depends on c only through a permutation of coordinates. Therefore the scalar group level CE objective within a fixed h group depends on the graph only through h, and the corresponding group optimal positive temperature is a function of h alone on $\mathcal{H}_+$.

This is an idealized population-concentration statement. It does not say that arbitrary graph-aware information is redundant in finite CSBM or in real graphs, where degree variation, off class composition noise, learned logits, or neighbor features may carry additional calibration information.

A.5 Entropy and Logit Taylor Expansions

For the binary expansion, let $\Delta = z_1 - z_2$. The binary predictive probability is $\sigma(\Delta)$, and the entropy is $H(\sigma(\Delta))$. A Taylor expansion around $\Delta = 0$ gives

$$H(\sigma(\Delta)) = \log 2 - \frac{\Delta^2}{8} + O(\Delta^4). \tag{33}$$

After normalizing by $\log 2$ and inverting,

$$|\Delta| = \sqrt{8 \log 2 (1 - e)} + O((1 - e)^{3/2}). \tag{34}$$

For the multi-class expansion, use softmax shift invariance and assume $\sum_k z_k = 0$. Around $z = 0$, the entropy satisfies

$$H(\mathrm{softmax}(z)) = \log K - \frac{\|z\|^2}{2K} + O(\|z\|^3). \tag{35}$$

The quadratic term is the Fisher information at the uniform distribution restricted to the centered subspace. Dividing by $\log K$ gives

$$1 - e = \frac{\|z\|^2}{2K \log K} + O(\|z\|^3). \tag{36}$$

Inversion yields

$$\|z^\circ\| = \sqrt{2K \log K (1 - e)} + O(1 - e). \tag{37}$$

For $K > 2$, the cubic term generally does not vanish, so the remainder is $O(1 - e)$ after taking the square root.

These expansions are local around the uniform predictive distribution. In HoTS, $\sqrt{2K \log K (1 - e_i)}$ is used as a simple monotone proxy for logit concentration over the full entropy range rather than as a global identity.

A.6 Binary CSBM Reliability Calculation

For $K = 2$, the centered simplex templates satisfy $m_2 = -m_1$. Under the centered class-template scores,

$$s_1 = \rho_0^{-1} m_1^\top (\bar{x} - \bar{\mu}), \qquad s_2 = -s_1. \tag{38}$$

Conditional on $Y = 1$ and $H = h$, the population-concentration model gives

$$s_1 \sim \mathcal{N}(\Gamma(h), v). \tag{39}$$

The class-template prediction is correct when $s_1 > s_2$, equivalently when $s_1 > 0$. Hence

$$q(h) = P(s_1 > 0 \mid Y = 1, H = h) = \Phi(\Gamma(h)/\sqrt{v}). \tag{40}$$

By symmetry, the same class-template correctness probability holds conditional on $Y = 2$.

Thus the exact logit reliability is

$$\operatorname{logit}(q(h)) = \operatorname{logit} \Phi(\Gamma(h)/\sqrt{v}). \tag{41}$$

The approximation

$$\operatorname{logit}(\Phi(x)) \approx \frac{\pi}{\sqrt{3}} x \tag{42}$$

is the standard variance matching logistic probit approximation. Combining it with the large $\bar{d}\tilde{h}$ relation $\Gamma(h) \approx \Gamma_0 \tilde{h}$ gives

$$\operatorname{logit}(q(h)) \approx \frac{\pi \Gamma_0}{\sqrt{3v}} \tilde{h}. \tag{43}$$

This linear relation is a moderate-margin approximation. It is not a large-margin asymptotic, since for very large positive x, $\operatorname{logit}(\Phi(x))$ is not exactly linear.

A.7 Proof of Proposition 2

The population regression slope of $\log T^*$ on $\hat{u}$ is

$$\frac{\operatorname{Cov}(\hat{u}, \log T^*)}{\operatorname{Var}(\hat{u})}. \tag{44}$$

Since $\log T^* = a - u$ and $\hat{u} = u + \eta$, with η independent of u and mean zero, we have

$$\operatorname{Cov}(\hat{u}, \log T^*) = -\operatorname{Var}(u) \tag{45}$$

and

$$\operatorname{Var}(\hat{u}) = \operatorname{Var}(u) + \tau^2. \tag{46}$$

Thus the fitted slope is

$$-\frac{\operatorname{Var}(u)}{\operatorname{Var}(u) + \tau^2}. \tag{47}$$

Because the model writes the slope as $-\alpha$, this gives

$$\alpha^* = \frac{\operatorname{Var}(u)}{\operatorname{Var}(u) + \tau^2}. \tag{48}$$

The strictness condition follows immediately.

A.8 Sparse-Overlap Gaussian Residual Approximation

This subsection records the technical approximation for neighbor-logits. It is used only to justify that, in a sparse-overlap regime, the Gaussian residual dependence among neighbor-logits is small after conditioning on the center feature x_i.

Condition on x_i. For two distinct neighbors $j, j' \in \mathcal{N}(i)$, define $S_{jj'}$ as the number of shared residual aggregation inputs in

$$(\mathcal{N}(j) \cup \{j\}) \cap (\mathcal{N}(j') \cup \{j'\}) \tag{49}$$

after excluding the common center node i. Suppose

$$\mathbb{E}[S_{jj'}] = O(K\bar{d}^2/n), \qquad \mathbb{E}[S_{jj'}^2] = O(K\bar{d}^2/n). \tag{50}$$

Work on the effective non-degenerate residual logit subspace, equivalently using the Moore-Penrose inverse for whitening.

Let W be the fixed linear head. Conditional on x_i, the residual covariance of one neighbor-logit vector is proportional to

$$\frac{\bar{d}\sigma^2 W W^\top}{(\bar{d}+1)^2}, \tag{51}$$

while the cross covariance between two neighbor residuals is proportional to

$$\frac{S_{jj'}\sigma^2 WW^\top}{(\bar{d}+1)^2}. \tag{52}$$

After whitening on the effective rank r logit subspace, the off-diagonal residual correlation block is

$$R_{jj'} = \frac{S_{jj'}}{\bar{d}} I_r, \qquad r \le K. \tag{53}$$

Thus

$$\|R_{jj'}\|_F^2 \le \frac{K S_{jj'}^2}{\bar{d}^2}. \tag{54}$$

By the sparse-overlap assumption, its expectation is $O(K^2/n)$.

Let $\mathcal{R}$ be the block off-diagonal whitened residual correlation matrix for all neighbor residual logits. The joint whitened Gaussian residual covariance is $I + \mathcal{R}$, whereas the product of marginals has covariance I. The Gaussian KL from the joint residual law to the product law is

$$-\frac{1}{2}\log\det(I+\mathcal{R}), \tag{55}$$

because $\mathcal{R}$ has zero diagonal blocks. If $\|\mathcal{R}\|_{\mathrm{op}} = o_p(1)$ and the log-determinant expansion is uniformly integrable, then the KL is

$$\frac{1}{4}\|\mathcal{R}\|_F^2 + o(\|\mathcal{R}\|_F^2) \tag{56}$$

in expectation. Since there are $O(\bar{d}^2)$ neighbor pairs and each contributes $O(K^2/n)$ in expectation, the expected Gaussian residual KL is

$$O(K^2\bar{d}^2/n). \tag{57}$$

Therefore, in the sparse-overlap regime $K^2\bar{d}^2/n \to 0$, the product approximation for neighbor-logits is leading-order accurate for the Gaussian residual dependence. This is not an exact full finite CSBM independence theorem. Shared class labels, finite degree composition randomness, and learned logit effects can introduce additional dependence terms unless separately controlled.

B Additional Empirical Results

This appendix provides additional empirical evidence supporting the main claims. Appendix B.1 reports complementary calibration metrics beyond ECE. Appendix B.2 breaks down the results by GCN and GAT backbones. Appendix B.3 records the additional backbone comparison and its experimental protocol. Appendix B.4 visualizes reliability diagrams and node-level confidence changes. Appendix B.5 checks the sparse-overlap KL scaling used in the theoretical approximation. Appendix B.6 evaluates calibration under random structural perturbations. Appendix B.7 examines whether oracle local homophily is associated with prediction correctness after conditioning on confidence. Appendix B.8 compares the residual reliability signal of oracle homophily with other structural statistics. Appendix B.9 compares HoTS with entropy-only scaling and examines learning versus fixing the homophily exponent.

B.1 Additional Calibration Metrics

ECE is the primary metric because it directly measures confidence and accuracy mismatch, but it is a binned statistic. Table 3 therefore reports two complementary summaries, mean NLL and degree-stratified ECE. The degree-stratified metric evaluates calibration after grouping nodes by degree, which is important for graph data because message passing reliability can vary with local connectivity.

Table 3: Pooled (GCN + GAT) mean NLL and degree-stratified ECE (%) across 18 datasets and 10 seeds per backbone.

Metric	TS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS (ours)
NLL	0.97	0.94	0.94	0.97	0.98	0.95	1.02	**0.92**
deg_ECE	6.50	6.32	6.13	**5.81**	6.45	6.16	7.16	5.90

B.2 Per-Backbone Results

The main table averages over GCN and GAT to summarize the benchmark level behavior. Tables 4, 5, and 6 give the corresponding per-backbone results. These tables are useful for diagnosing where the aggregate trends come from. HoTS is not uniformly best for every backbone and dataset, but its average behavior remains strong across the two architectures.

Table 4: Per-backbone ECE (%, mean±std over 10 seeds per backbone). Best per (dataset, backbone) row in red, second in **bold**, third underlined (excluding Uncal).

Dataset	Backbone	Uncal	TS	VS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS
Homophilic											
Cora	GCN	21.49 ± 1.15	3.42 ± 0.70	**3.05 ± 0.97**	3.19 ± 0.77	3.66 ± 0.45	4.08 ± 0.88	3.30 ± 0.96	4.16 ± 0.92	3.20 ± 0.69	2.28 ± 0.76
	GAT	17.79 ± 2.37	2.86 ± 0.83	2.75 ± 0.48	**2.69 ± 0.78**	3.49 ± 0.64	4.17 ± 0.67	2.68 ± 0.85	3.55 ± 0.56	2.95 ± 0.75	2.72 ± 0.59
CiteSeer	GCN	21.87 ± 3.46	**2.91 ± 0.42**	3.16 ± 0.73	2.90 ± 0.44	3.68 ± 0.98	4.93 ± 1.08	3.42 ± 0.77	5.66 ± 1.10	3.01 ± 0.47	3.36 ± 1.02
	GAT	15.26 ± 1.17	3.17 ± 0.74	3.24 ± 0.32	2.96 ± 0.78	3.88 ± 0.71	4.69 ± 0.22	2.96 ± 0.87	4.49 ± 0.63	3.18 ± 0.65	3.65 ± 0.44
PubMed	GCN	13.90 ± 1.58	1.18 ± 0.20	1.44 ± 0.31	1.18 ± 0.23	1.45 ± 0.65	1.28 ± 1.31	**0.91 ± 0.30**	1.14 ± 0.38	0.74 ± 0.24	1.25 ± 0.37
	GAT	10.41 ± 0.83	0.89 ± 0.28	0.87 ± 0.23	0.83 ± 0.28	1.00 ± 0.57	0.94 ± 0.41	**0.85 ± 0.25**	1.13 ± 0.39	0.92 ± 0.33	1.27 ± 0.22
CoraFull	GCN	28.50 ± 0.92	3.40 ± 0.44	3.22 ± 0.58	2.71 ± 0.39	3.94 ± 0.52	2.87 ± 0.54	3.38 ± 0.41	3.74 ± 0.84	2.43 ± 0.62	**2.47 ± 0.55**
	GAT	37.35 ± 0.88	**2.11 ± 0.52**	2.47 ± 0.42	1.68 ± 0.31	3.94 ± 0.73	2.26 ± 0.52	2.17 ± 0.61	3.71 ± 1.10	5.27 ± 1.01	2.99 ± 0.93
Computers	GCN	5.12 ± 0.81	2.33 ± 0.48	2.27 ± 0.35	2.26 ± 0.58	1.51 ± 0.35	1.20 ± 0.24	1.80 ± 0.59	2.62 ± 0.58	**1.23 ± 0.35**	1.62 ± 0.40
	GAT	6.60 ± 1.52	2.04 ± 0.37	**1.87 ± 0.45**	1.98 ± 0.36	2.06 ± 0.57	1.57 ± 0.33	2.01 ± 0.33	2.70 ± 0.86	2.01 ± 0.53	1.90 ± 0.52
Photo	GCN	3.68 ± 0.62	1.24 ± 0.32	1.30 ± 0.23	1.24 ± 0.34	0.96 ± 0.37	1.29 ± 0.47	1.17 ± 0.25	1.69 ± 0.35	1.18 ± 0.41	**0.98 ± 0.39**
	GAT	4.93 ± 0.69	**1.22 ± 0.30**	1.24 ± 0.27	1.37 ± 0.59	1.31 ± 0.52	1.41 ± 0.36	1.32 ± 0.37	1.16 ± 0.29	1.28 ± 0.36	1.44 ± 0.34
CS	GCN	1.58 ± 0.19	1.57 ± 0.18	1.44 ± 0.18	1.49 ± 0.18	**1.37 ± 0.36**	2.79 ± 0.83	1.57 ± 0.22	1.59 ± 0.16	3.53 ± 0.43	1.08 ± 0.22
	GAT	3.86 ± 0.51	2.95 ± 0.64	2.62 ± 0.46	2.95 ± 0.65	1.32 ± 0.26	1.99 ± 0.61	2.66 ± 0.51	3.69 ± 3.70	2.98 ± 0.60	**1.94 ± 0.55**
Physics	GCN	0.65 ± 0.16	0.62 ± 0.14	0.64 ± 0.15	0.55 ± 0.11	0.39 ± 0.13	0.80 ± 0.32	0.47 ± 0.11	0.64 ± 0.24	1.80 ± 0.28	**0.41 ± 0.13**
	GAT	1.75 ± 0.38	1.12 ± 0.15	1.05 ± 0.14	1.09 ± 0.19	**0.66 ± 0.16**	0.64 ± 0.14	0.97 ± 0.19	1.37 ± 0.71	1.03 ± 0.25	0.67 ± 0.23
Heterophilic											
Texas	GCN	25.38 ± 6.99	25.08 ± 7.10	19.96 ± 7.05	24.16 ± 6.21	**19.30 ± 7.09**	23.18 ± 6.36	21.83 ± 7.24	20.14 ± 4.94	27.84 ± 8.25	16.67 ± 4.49
	GAT	18.35 ± 7.91	18.19 ± 8.36	16.50 ± 6.35	14.38 ± 7.43	16.08 ± 5.88	17.47 ± 8.10	13.87 ± 4.94	16.43 ± 4.28	21.72 ± 6.86	**14.08 ± 6.79**
Cornell	GCN	22.49 ± 9.37	21.70 ± 8.67	22.18 ± 5.77	19.19 ± 8.74	19.58 ± 7.83	**16.61 ± 6.67**	20.04 ± 8.45	21.97 ± 7.92	24.96 ± 7.76	13.31 ± 5.35
	GAT	17.76 ± 5.75	18.40 ± 5.37	17.11 ± 6.36	17.16 ± 6.08	16.53 ± 6.57	**15.26 ± 7.26**	17.11 ± 7.03	17.19 ± 7.12	25.38 ± 5.83	15.24 ± 6.11
Wisconsin	GCN	21.44 ± 7.41	21.40 ± 6.80	22.26 ± 6.65	**18.03 ± 6.18**	20.66 ± 6.40	20.43 ± 6.08	20.69 ± 5.83	22.12 ± 6.92	26.29 ± 7.65	16.59 ± 5.09
	GAT	17.18 ± 6.55	16.72 ± 6.79	16.97 ± 6.72	15.60 ± 7.18	15.60 ± 6.16	17.11 ± 6.99	**14.96 ± 6.29**	17.75 ± 6.55	22.85 ± 6.59	12.75 ± 4.96
Chameleon	GCN	12.73 ± 2.15	12.81 ± 2.18	12.05 ± 2.81	13.03 ± 1.92	9.34 ± 2.08	10.39 ± 2.03	12.25 ± 2.08	13.28 ± 2.41	12.80 ± 2.25	**9.44 ± 2.76**
	GAT	10.27 ± 2.23	10.34 ± 2.31	10.43 ± 2.50	9.36 ± 4.00	**8.56 ± 1.34**	8.01 ± 2.08	9.08 ± 3.05	8.72 ± 2.71	13.55 ± 1.47	8.67 ± 1.78
Squirrel	GCN	8.09 ± 1.53	8.10 ± 1.45	8.19 ± 1.58	8.57 ± 1.14	5.13 ± 1.98	**5.14 ± 1.55**	8.22 ± 1.53	10.11 ± 1.33	7.48 ± 1.44	5.69 ± 1.22
	GAT	4.29 ± 1.49	4.29 ± 1.53	4.15 ± 1.48	4.16 ± 1.61	3.81 ± 1.78	**3.62 ± 1.25**	3.52 ± 1.26	4.50 ± 0.98	5.61 ± 2.26	4.24 ± 1.46
Actor	GCN	2.23 ± 0.85	2.21 ± 0.87	1.88 ± 0.60	**2.11 ± 0.93**	2.33 ± 1.04	2.16 ± 0.94	**2.11 ± 1.06**	2.97 ± 1.53	5.37 ± 2.17	2.33 ± 1.06
	GAT	2.71 ± 1.04	2.73 ± 0.73	2.04 ± 0.88	2.82 ± 0.77	3.04 ± 0.84	**2.57 ± 0.74**	2.92 ± 0.76	5.79 ± 1.65	4.18 ± 1.33	2.65 ± 0.53
Roman-Empire	GCN	12.44 ± 0.80	2.39 ± 0.55	1.57 ± 0.53	**2.32 ± 0.58**	5.02 ± 0.65	2.42 ± 0.55	3.05 ± 0.58	2.76 ± 0.60	3.21 ± 1.23	4.45 ± 0.59
	GAT	12.10 ± 1.29	2.89 ± 0.66	**2.64 ± 0.73**	3.12 ± 0.73	3.10 ± 0.80	3.17 ± 0.91	2.63 ± 0.64	3.99 ± 0.82	2.66 ± 0.71	4.26 ± 0.93
tolokers	GCN	3.72 ± 0.76	3.77 ± 0.81	3.75 ± 0.81	3.80 ± 0.77	3.40 ± 1.00	**3.24 ± 0.69**	3.33 ± 1.10	2.29 ± 0.68	4.13 ± 0.75	3.52 ± 0.81
	GAT	4.88 ± 1.06	4.67 ± 0.60	3.57 ± 0.79	4.65 ± 0.61	2.45 ± 0.58	**2.09 ± 0.42**	4.18 ± 0.42	1.79 ± 0.45	4.38 ± 0.50	2.98 ± 0.65
Large											
ogbn-arxiv	GCN	3.12 ± 0.24	3.14 ± 0.22	3.03 ± 0.13	3.08 ± 0.23	**1.19 ± 0.13**	2.02 ± 0.37	3.37 ± 0.17	5.09 ± 0.56	2.53 ± 0.64	1.15 ± 0.11
	GAT	7.01 ± 0.52	2.16 ± 0.13	2.41 ± 0.13	2.06 ± 0.17	1.40 ± 0.15	**1.62 ± 0.20**	1.89 ± 0.25	3.92 ± 0.65	1.85 ± 0.17	1.91 ± 0.17
Reddit	GCN	6.93 ± 0.14	2.35 ± 0.10	2.39 ± 0.08	2.22 ± 0.17	1.19 ± 0.06	1.41 ± 0.10	–	3.22 ± 0.18	1.69 ± 0.18	**1.35 ± 0.13**
	GAT	3.77 ± 0.20	1.02 ± 0.12	1.28 ± 0.20	1.24 ± 0.19	1.09 ± 0.16	0.48 ± 0.10	–	**0.61 ± 0.08**	0.80 ± 0.10	1.15 ± 0.17

Table 5: Per-backbone degree-stratified ECE (%, mean±std). Best per (dataset, backbone) row in red, second in **bold**, third underlined (excluding Uncal).

Dataset	Backbone	Uncal	TS	VS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS
Homophilic											
Cora	GCN	21.47 ± 1.15	2.79 ± 0.53	2.72 ± 0.82	2.85 ± 0.58	**2.70 ± 0.43**	3.10 ± 0.51	2.89 ± 0.57	3.23 ± 0.54	2.83 ± 0.54	2.59 ± 0.41
	GAT	17.79 ± 2.37	**2.17 ± 0.71**	2.09 ± 0.73	2.35 ± 0.75	2.30 ± 0.51	2.72 ± 0.73	**2.17 ± 0.79**	2.53 ± 0.48	2.25 ± 0.57	**2.17 ± 0.59**
CiteSeer	GCN	21.86 ± 3.45	2.71 ± 0.43	2.76 ± 0.59	2.71 ± 0.55	3.23 ± 0.80	3.32 ± 0.54	3.15 ± 0.75	4.04 ± 0.91	2.85 ± 0.54	2.91 ± 0.93
	GAT	15.24 ± 1.18	**2.58 ± 0.49**	2.83 ± 0.52	2.46 ± 0.39	3.36 ± 0.76	2.77 ± 0.48	2.94 ± 0.87	3.04 ± 0.88	2.67 ± 0.39	3.13 ± 0.45
PubMed	GCN	13.94 ± 1.52	2.58 ± 0.34	2.92 ± 0.34	2.72 ± 0.33	**1.83 ± 0.36**	1.84 ± 1.47	2.19 ± 0.39	2.16 ± 0.57	0.92 ± 0.18	1.98 ± 0.25
	GAT	10.40 ± 0.83	0.82 ± 0.26	0.75 ± 0.13	**0.79 ± 0.23**	0.80 ± 0.19	0.96 ± 0.49	0.80 ± 0.29	1.02 ± 0.38	0.87 ± 0.32	0.84 ± 0.29
CoraFull	GCN	28.50 ± 0.92	4.83 ± 0.46	3.24 ± 0.60	4.89 ± 0.46	2.64 ± 0.64	3.75 ± 0.44	4.89 ± 0.46	3.48 ± 0.58	**2.33 ± 0.55**	1.56 ± 0.47
	GAT	37.35 ± 0.88	**1.81 ± 0.57**	2.36 ± 0.24	1.42 ± 0.36	3.06 ± 0.97	2.75 ± 0.89	1.94 ± 0.64	2.30 ± 1.30	5.23 ± 1.03	2.15 ± 0.96
Computers	GCN	6.33 ± 0.45	4.02 ± 0.52	3.11 ± 0.39	4.06 ± 0.52	2.19 ± 0.44	**1.82 ± 0.37**	3.23 ± 0.95	3.36 ± 0.64	1.40 ± 0.32	2.76 ± 0.48
	GAT	6.59 ± 1.52	2.19 ± 0.27	2.62 ± 0.49	2.19 ± 0.25	2.14 ± 0.28	2.27 ± 0.31	**2.12 ± 0.29**	2.83 ± 0.94	2.34 ± 0.28	2.09 ± 0.27
Photo	GCN	4.23 ± 0.60	1.97 ± 0.28	1.89 ± 0.35	1.99 ± 0.31	**1.46 ± 0.44**	1.71 ± 0.49	1.94 ± 0.25	2.22 ± 0.37	1.37 ± 0.26	1.47 ± 0.38
	GAT	4.88 ± 0.72	1.67 ± 0.41	1.74 ± 0.28	1.86 ± 0.60	1.82 ± 0.47	**1.49 ± 0.33**	1.72 ± 0.46	1.35 ± 0.22	1.55 ± 0.34	1.72 ± 0.40
CS	GCN	1.72 ± 0.21	1.67 ± 0.20	1.66 ± 0.17	1.64 ± 0.19	**1.40 ± 0.31**	2.52 ± 0.78	1.72 ± 0.19	1.66 ± 0.16	3.45 ± 0.49	1.15 ± 0.20
	GAT	3.58 ± 0.73	2.45 ± 0.70	2.25 ± 0.52	2.51 ± 0.81	**2.15 ± 0.37**	1.19 ± 0.35	**2.15 ± 0.54**	3.19 ± 3.17	2.56 ± 0.78	2.52 ± 0.59
Physics	GCN	0.98 ± 0.10	0.98 ± 0.12	0.97 ± 0.11	0.93 ± 0.15	**0.80 ± 0.09**	0.93 ± 0.40	0.99 ± 0.11	1.00 ± 0.14	1.77 ± 0.31	0.79 ± 0.09
	GAT	1.72 ± 0.41	0.80 ± 0.26	0.95 ± 0.24	0.90 ± 0.27	0.72 ± 0.12	0.45 ± 0.19	0.74 ± 0.27	0.76 ± 0.57	**0.67 ± 0.27**	0.68 ± 0.20
Heterophilic											
Texas	GCN	27.24 ± 7.36	27.23 ± 7.36	25.64 ± 5.45	26.84 ± 7.54	**24.57 ± 7.08**	25.94 ± 7.40	25.66 ± 7.36	25.22 ± 3.60	28.26 ± 7.01	23.07 ± 6.59
	GAT	22.51 ± 6.55	22.41 ± 6.43	22.33 ± 5.74	20.57 ± 5.13	21.64 ± 5.84	21.53 ± 7.23	**20.29 ± 5.81**	19.99 ± 2.65	23.93 ± 5.91	21.86 ± 5.44
Cornell	GCN	22.47 ± 8.53	21.50 ± 7.66	20.96 ± 7.38	19.38 ± 7.61	20.76 ± 7.61	**18.06 ± 7.45**	19.84 ± 6.95	18.34 ± 7.05	23.67 ± 6.87	17.56 ± 6.45
	GAT	19.97 ± 6.63	20.04 ± 6.68	19.84 ± 6.57	19.40 ± 6.37	18.57 ± 7.09	16.89 ± 6.27	18.44 ± 6.95	18.49 ± 4.90	24.89 ± 5.57	**17.98 ± 6.62**
Wisconsin	GCN	19.96 ± 7.01	20.10 ± 6.68	21.57 ± 4.96	**17.37 ± 5.35**	18.31 ± 5.21	19.37 ± 5.98	18.37 ± 5.07	18.95 ± 5.23	24.29 ± 8.28	16.52 ± 4.10
	GAT	18.88 ± 6.28	18.81 ± 6.27	19.18 ± 6.14	17.93 ± 6.29	17.28 ± 5.32	**16.99 ± 6.08**	17.22 ± 5.15	18.17 ± 7.40	23.67 ± 5.96	15.96 ± 3.34
Chameleon	GCN	12.07 ± 2.00	12.07 ± 2.00	11.98 ± 2.08	12.09 ± 1.93	11.88 ± 1.59	10.60 ± 1.47	11.79 ± 1.84	**11.75 ± 2.21**	13.62 ± 2.16	12.22 ± 1.62
	GAT	12.56 ± 2.07	12.52 ± 2.09	12.33 ± 2.01	12.56 ± 2.61	12.29 ± 2.15	**10.91 ± 2.10**	12.21 ± 2.38	10.88 ± 2.01	15.28 ± 2.80	12.74 ± 2.14
Squirrel	GCN	6.27 ± 1.20	6.33 ± 1.20	6.42 ± 1.00	7.65 ± 0.99	**5.45 ± 1.09**	5.51 ± 1.08	7.09 ± 1.09	5.36 ± 1.03	6.96 ± 1.97	5.78 ± 1.12
	GAT	6.48 ± 1.97	6.50 ± 1.97	6.78 ± 1.63	6.70 ± 2.05	6.35 ± 2.10	**6.32 ± 2.10**	6.41 ± 2.07	5.78 ± 1.42	7.06 ± 2.21	6.34 ± 2.15
Actor	GCN	2.36 ± 0.69	2.44 ± 0.65	2.21 ± 0.52	2.37 ± 0.68	2.49 ± 0.74	**2.25 ± 0.73**	2.41 ± 0.68	2.96 ± 1.38	5.41 ± 2.05	2.27 ± 0.72
	GAT	1.78 ± 0.35	1.87 ± 0.33	1.62 ± 0.48	1.88 ± 0.27	1.93 ± 0.31	1.91 ± 0.42	1.88 ± 0.34	4.03 ± 1.85	3.64 ± 1.07	**1.79 ± 0.40**
Roman-Empire	GCN	12.57 ± 0.67	**3.12 ± 0.44**	4.48 ± 0.51	3.15 ± 0.51	3.40 ± 0.49	3.01 ± 0.47	3.32 ± 0.53	4.15 ± 0.56	3.65 ± 0.85	3.74 ± 0.63
	GAT	12.08 ± 1.28	6.96 ± 0.35	6.12 ± 0.52	7.04 ± 0.34	6.88 ± 0.34	**5.42 ± 0.28**	6.68 ± 0.31	2.45 ± 0.67	6.70 ± 0.28	7.08 ± 0.37
tolokers	GCN	2.26 ± 0.23	2.31 ± 0.36	2.34 ± 0.27	**2.30 ± 0.35**	2.33 ± 0.28	3.19 ± 0.67	2.51 ± 0.50	2.75 ± 0.57	2.24 ± 0.37	2.47 ± 0.21
	GAT	4.22 ± 0.82	3.62 ± 0.29	4.00 ± 0.42	3.62 ± 0.29	3.65 ± 0.30	2.17 ± 0.22	3.68 ± 0.30	**2.56 ± 0.33**	2.83 ± 0.35	3.60 ± 0.43
Large											
ogbn-arxiv	GCN	1.67 ± 0.20	1.71 ± 0.18	1.62 ± 0.12	1.71 ± 0.18	**1.44 ± 0.08**	1.37 ± 0.28	2.50 ± 0.26	4.59 ± 0.64	1.93 ± 0.57	1.74 ± 0.12
	GAT	6.95 ± 0.46	3.31 ± 0.17	3.86 ± 0.12	3.29 ± 0.16	4.11 ± 0.15	2.02 ± 0.15	3.30 ± 0.17	3.38 ± 0.75	**2.35 ± 0.24**	4.28 ± 0.15
Reddit	GCN	6.93 ± 0.14	1.86 ± 0.10	2.10 ± 0.12	1.91 ± 0.10	1.21 ± 0.06	1.52 ± 0.11	–	2.31 ± 0.26	1.41 ± 0.14	**1.33 ± 0.12**
	GAT	4.94 ± 0.18	3.24 ± 0.10	3.38 ± 0.10	3.44 ± 0.10	3.46 ± 0.11	0.67 ± 0.09	–	1.16 ± 0.15	**0.83 ± 0.16**	3.53 ± 0.12

Table 6: Per-backbone NLL (mean±std). Best per (dataset, backbone) row in red, second in **bold**, third underlined (excluding Uncal).

Dataset	Backbone	Uncal	TS	VS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS
Homophilic											
Cora	GCN	0.681 ± 0.025	0.518 ± 0.031	0.481 ± 0.026	0.512 ± 0.030	0.503 ± 0.028	0.572 ± 0.053	0.518 ± 0.030	0.528 ± 0.032	0.516 ± 0.030	**0.497 ± 0.028**
	GAT	0.636 ± 0.036	0.516 ± 0.027	**0.506 ± 0.022**	0.515 ± 0.026	0.513 ± 0.024	0.553 ± 0.026	0.520 ± 0.026	0.523 ± 0.021	0.516 ± 0.027	0.504 ± 0.023
CiteSeer	GCN	0.977 ± 0.066	0.849 ± 0.046	0.809 ± 0.023	**0.844 ± 0.042**	0.853 ± 0.052	0.878 ± 0.042	0.853 ± 0.046	0.884 ± 0.039	0.849 ± 0.048	0.859 ± 0.050
	GAT	0.878 ± 0.009	0.801 ± 0.012	0.790 ± 0.011	**0.800 ± 0.012**	0.803 ± 0.013	0.832 ± 0.017	0.807 ± 0.011	0.832 ± 0.016	0.801 ± 0.012	0.805 ± 0.013
PubMed	GCN	0.459 ± 0.017	0.380 ± 0.005	0.377 ± 0.005	0.377 ± 0.005	0.376 ± 0.006	**0.368 ± 0.017**	0.375 ± 0.006	0.371 ± 0.005	0.365 ± 0.005	0.376 ± 0.005
	GAT	0.430 ± 0.005	0.378 ± 0.006	**0.375 ± 0.005**	0.378 ± 0.006	0.379 ± 0.006	**0.375 ± 0.005**	0.379 ± 0.006	0.373 ± 0.005	0.379 ± 0.006	0.378 ± 0.005
CoraFull	GCN	1.947 ± 0.016	1.590 ± 0.030	1.324 ± 0.017	1.582 ± 0.029	1.572 ± 0.029	1.563 ± 0.030	1.590 ± 0.030	**1.395 ± 0.034**	1.562 ± 0.030	1.569 ± 0.028
	GAT	2.256 ± 0.022	1.667 ± 0.026	1.258 ± 0.023	1.662 ± 0.026	1.684 ± 0.028	1.648 ± 0.035	1.668 ± 0.026	**1.439 ± 0.039**	1.689 ± 0.029	1.673 ± 0.028
Computers	GCN	0.400 ± 0.012	0.389 ± 0.015	0.367 ± 0.011	0.381 ± 0.015	0.373 ± 0.014	**0.363 ± 0.013**	0.381 ± 0.020	0.375 ± 0.007	0.360 ± 0.013	0.371 ± 0.017
	GAT	0.480 ± 0.014	0.455 ± 0.020	0.400 ± 0.007	0.453 ± 0.021	0.452 ± 0.019	0.442 ± 0.017	0.457 ± 0.021	**0.406 ± 0.016**	0.453 ± 0.020	0.450 ± 0.020
Photo	GCN	0.251 ± 0.010	0.242 ± 0.013	0.235 ± 0.010	0.239 ± 0.013	0.240 ± 0.012	0.246 ± 0.015	0.242 ± 0.013	0.244 ± 0.011	0.239 ± 0.018	**0.238 ± 0.009**
	GAT	0.308 ± 0.015	0.297 ± 0.018	**0.278 ± 0.014**	0.285 ± 0.015	0.297 ± 0.018	0.291 ± 0.020	0.302 ± 0.019	0.270 ± 0.013	0.294 ± 0.026	0.297 ± 0.018
CS	GCN	0.275 ± 0.008	0.276 ± 0.008	0.275 ± 0.008	0.275 ± 0.008	0.268 ± 0.008	0.340 ± 0.048	0.276 ± 0.009	0.278 ± 0.008	0.308 ± 0.017	0.268 ± 0.008
	GAT	0.385 ± 0.011	0.384 ± 0.014	0.375 ± 0.011	0.382 ± 0.015	0.368 ± 0.013	0.368 ± 0.027	0.386 ± 0.015	0.421 ± 0.145	0.414 ± 0.026	0.368 ± 0.013
Physics	GCN	0.140 ± 0.004	0.140 ± 0.004	0.139 ± 0.004	0.139 ± 0.004	0.138 ± 0.004	0.149 ± 0.015	0.139 ± 0.004	0.141 ± 0.007	0.153 ± 0.007	0.138 ± 0.004
	GAT	0.184 ± 0.003	0.182 ± 0.004	0.177 ± 0.003	0.178 ± 0.005	**0.176 ± 0.004**	0.167 ± 0.003	0.181 ± 0.005	0.192 ± 0.051	0.184 ± 0.008	0.177 ± 0.004
Heterophilic											
Texas	GCN	1.662 ± 0.354	1.658 ± 0.349	1.566 ± 0.396	1.517 ± 0.145	1.517 ± 0.231	1.587 ± 0.179	**1.514 ± 0.166**	1.701 ± 0.279	1.911 ± 0.532	1.402 ± 0.088
	GAT	1.452 ± 0.266	1.443 ± 0.261	1.396 ± 0.223	1.286 ± 0.174	1.301 ± 0.138	1.581 ± 0.745	**1.272 ± 0.099**	1.368 ± 0.234	1.644 ± 0.350	1.227 ± 0.083
Cornell	GCN	1.707 ± 0.169	1.687 ± 0.148	1.633 ± 0.175	**1.563 ± 0.068**	1.595 ± 0.080	1.711 ± 0.309	1.653 ± 0.218	2.012 ± 0.520	1.864 ± 0.275	1.514 ± 0.089
	GAT	1.595 ± 0.401	1.595 ± 0.392	1.580 ± 0.400	**1.459 ± 0.139**	1.532 ± 0.235	1.638 ± 0.559	1.532 ± 0.182	1.624 ± 0.273	1.811 ± 0.580	1.426 ± 0.087
Wisconsin	GCN	1.674 ± 0.317	1.673 ± 0.309	1.617 ± 0.328	**1.508 ± 0.079**	1.537 ± 0.163	1.768 ± 0.480	1.580 ± 0.170	1.616 ± 0.185	2.026 ± 0.668	1.454 ± 0.082
	GAT	1.656 ± 0.597	1.649 ± 0.589	1.644 ± 0.590	1.527 ± 0.610	1.480 ± 0.341	1.606 ± 0.478	**1.473 ± 0.382**	1.667 ± 0.579	2.114 ± 1.071	1.321 ± 0.041
Chameleon	GCN	1.369 ± 0.023	1.368 ± 0.023	1.368 ± 0.020	1.333 ± 0.021	**1.327 ± 0.035**	1.333 ± 0.056	1.349 ± 0.030	1.457 ± 0.164	1.480 ± 0.063	1.317 ± 0.029
	GAT	1.270 ± 0.031	1.268 ± 0.031	1.268 ± 0.031	**1.206 ± 0.063**	1.221 ± 0.044	1.198 ± 0.045	1.221 ± 0.057	1.238 ± 0.065	1.421 ± 0.087	1.216 ± 0.048
Squirrel	GCN	1.474 ± 0.014	1.474 ± 0.014	1.472 ± 0.014	1.456 ± 0.020	**1.442 ± 0.014**	1.432 ± 0.013	1.460 ± 0.014	1.716 ± 0.132	1.526 ± 0.035	1.445 ± 0.015
	GAT	1.536 ± 0.016	1.536 ± 0.016	1.536 ± 0.016	1.527 ± 0.019	1.531 ± 0.016	1.532 ± 0.015	1.531 ± 0.017	1.565 ± 0.030	1.564 ± 0.022	**1.529 ± 0.019**
Actor	GCN	1.548 ± 0.006	1.548 ± 0.006	**1.547 ± 0.005**	1.548 ± 0.006	1.548 ± 0.006	1.548 ± 0.007	1.549 ± 0.006	1.542 ± 0.011	1.564 ± 0.012	1.551 ± 0.006
	GAT	1.557 ± 0.004	1.557 ± 0.004	**1.547 ± 0.005**	1.556 ± 0.004	1.556 ± 0.004	1.555 ± 0.003	1.556 ± 0.004	1.524 ± 0.009	1.567 ± 0.014	1.557 ± 0.004
Roman-Empire	GCN	1.627 ± 0.024	1.559 ± 0.020	**1.494 ± 0.016**	1.559 ± 0.020	1.567 ± 0.021	1.544 ± 0.020	1.566 ± 0.019	1.455 ± 0.037	1.562 ± 0.019	1.565 ± 0.020
	GAT	2.019 ± 0.032	1.969 ± 0.035	**1.898 ± 0.039**	1.968 ± 0.035	1.972 ± 0.035	1.981 ± 0.033	1.968 ± 0.035	1.463 ± 0.039	1.974 ± 0.037	1.972 ± 0.035
tolokers	GCN	0.469 ± 0.005	0.468 ± 0.005	0.468 ± 0.005	0.468 ± 0.005	0.467 ± 0.005	**0.460 ± 0.005**	0.464 ± 0.004	0.432 ± 0.011	0.466 ± 0.005	0.469 ± 0.003
	GAT	0.492 ± 0.005	0.490 ± 0.004	0.484 ± 0.004	0.490 ± 0.004	0.482 ± 0.005	**0.465 ± 0.003**	0.490 ± 0.004	0.429 ± 0.013	0.483 ± 0.004	0.482 ± 0.003
Large											
ogbn-arxiv	GCN	0.987 ± 0.005	0.987 ± 0.005	0.980 ± 0.003	0.984 ± 0.005	0.974 ± 0.004	0.962 ± 0.011	0.991 ± 0.005	0.999 ± 0.008	**0.971 ± 0.012**	0.972 ± 0.004
	GAT	1.239 ± 0.003	1.217 ± 0.006	1.191 ± 0.003	1.216 ± 0.006	1.209 ± 0.007	**1.202 ± 0.006**	1.213 ± 0.006	1.210 ± 0.005	1.210 ± 0.006	1.210 ± 0.007
Reddit	GCN	0.355 ± 0.003	0.321 ± 0.004	0.315 ± 0.004	0.314 ± 0.007	**0.301 ± 0.004**	0.299 ± 0.004	–	0.346 ± 0.004	0.306 ± 0.008	**0.301 ± 0.004**
	GAT	0.313 ± 0.007	0.306 ± 0.007	0.303 ± 0.006	0.296 ± 0.006	0.303 ± 0.007	**0.270 ± 0.006**	–	0.256 ± 0.004	0.273 ± 0.006	0.304 ± 0.008

B.3 Calibration Across GNN Backbones

Table 7 extends the GCN and GAT evaluations in Appendix B.2 to GCNII, GPR-GNN, GIN, and GraphSAGE. Each backbone is evaluated on the same 17 datasets and 10 seeds under the random train/validation/test split protocol, using the same calibration baselines as the main comparison. Within each dataset, backbone, and seed, all calibrators use the same frozen classifier logits. HoTS uses the observed-label-only predictor described in Appendix C.5. Reddit is excluded because some of the compared pipelines exceed the available memory. The table reports the macro-average of the per-dataset mean ECE values.

Table 7: Calibration across additional GNN backbones. Mean ECE (%) over 17 datasets, with 10 seeds per dataset. Reddit is excluded. Lower is better; best per row in **bold**.

Backbone	TS	VS	ETS	HTS	CaGCN	GATS	GETS	WATS	HoTS
GCNII	6.82	6.08	6.29	6.23	**5.79**	6.35	6.79	7.74	5.84
GPR-GNN	6.85	5.95	6.47	5.73	6.05	5.87	7.11	7.78	**5.49**
GIN	6.32	5.67	5.37	5.23	6.37	5.31	6.33	8.98	**4.97**
GraphSAGE	6.60	6.10	5.53	5.75	6.63	5.79	6.51	9.20	**5.11**

HoTS achieves the lowest mean ECE on GPR-GNN, GIN, and GraphSAGE, and ranks second on GCNII. These results support its applicability across different GNN architectures.

B.4 Reliability Diagrams and Node-Level Diagnostics

Figure 2 shows a representative reliability diagram on Reddit (GCN backbone). HoTS reduces the visible calibration gap left after standard temperature scaling. Pooled ECE decreases from 6.93% before calibration to 1.34% after calibration, compared with 2.34% for TS. Figure 3 extends this view across all benchmarks.

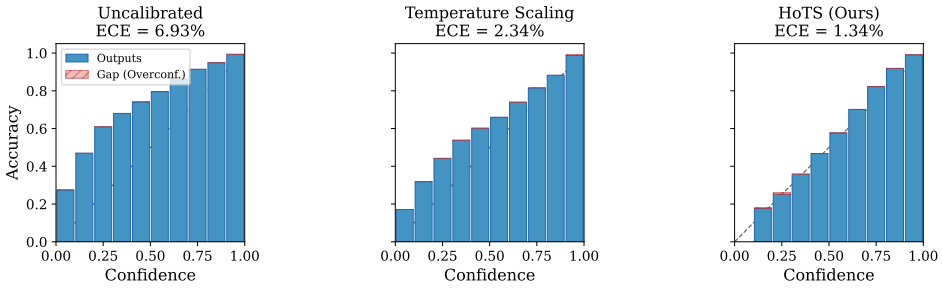

Figure 2: Reliability diagrams on Reddit with a GCN backbone, pooling test nodes across 10 seeds into 10 equal-width confidence bins. TS reduces the calibration gap, and HoTS further improves calibration by allowing the temperature to vary across nodes.

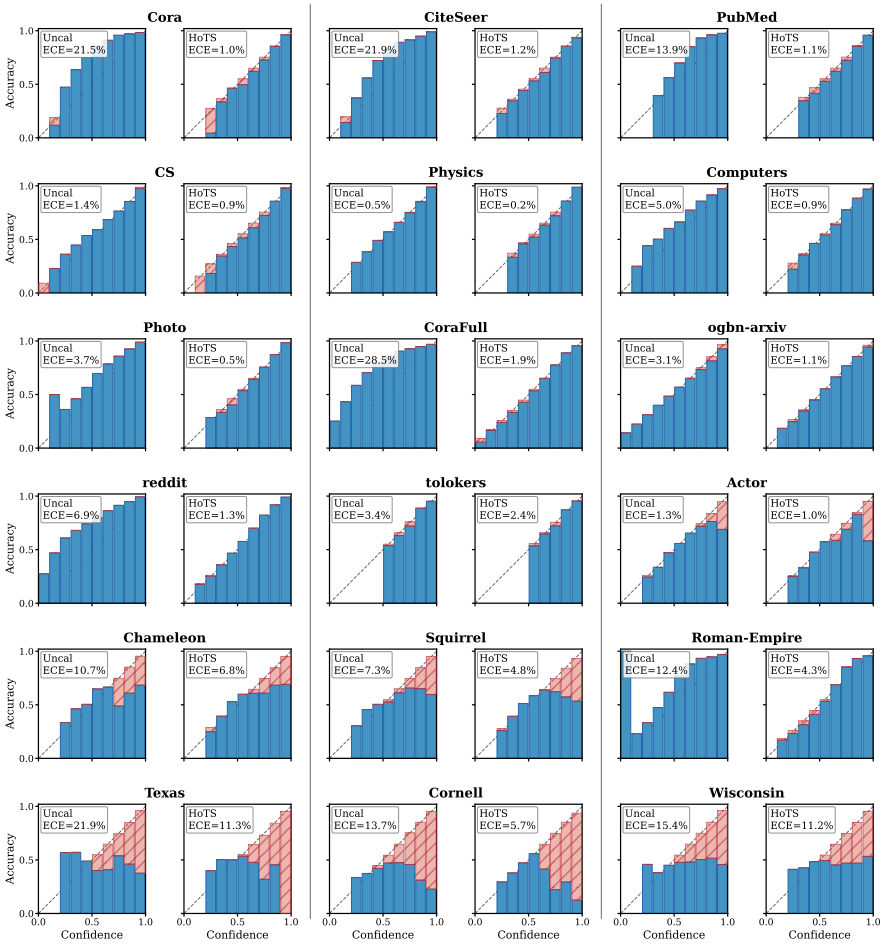

Figure 3: Reliability diagrams across the 18 benchmark datasets. Each pair compares the uncalibrated GCN and HoTS, pooling test nodes across 10 seeds into 10 equal-width confidence bins. The figure is intended as a diagnostic complement to Table 1; pooled ECE differs from the mean per-seed ECE reported in the tables.

Figure 4 visualizes node-level confidence gaps on representative homophilic and heterophilic graphs. The horizontal axis evaluates the homophily-dependent temperature correction using oracle homophily, and the vertical axis is predicted confidence minus the node's correctness indicator. Coloring by oracle homophily and by entropy complement relates these gaps to neighborhood composition and raw logit concentration. Oracle homophily is computed from true labels for this retrospective diagnostic; the implemented HoTS temperature map uses estimated homophily.

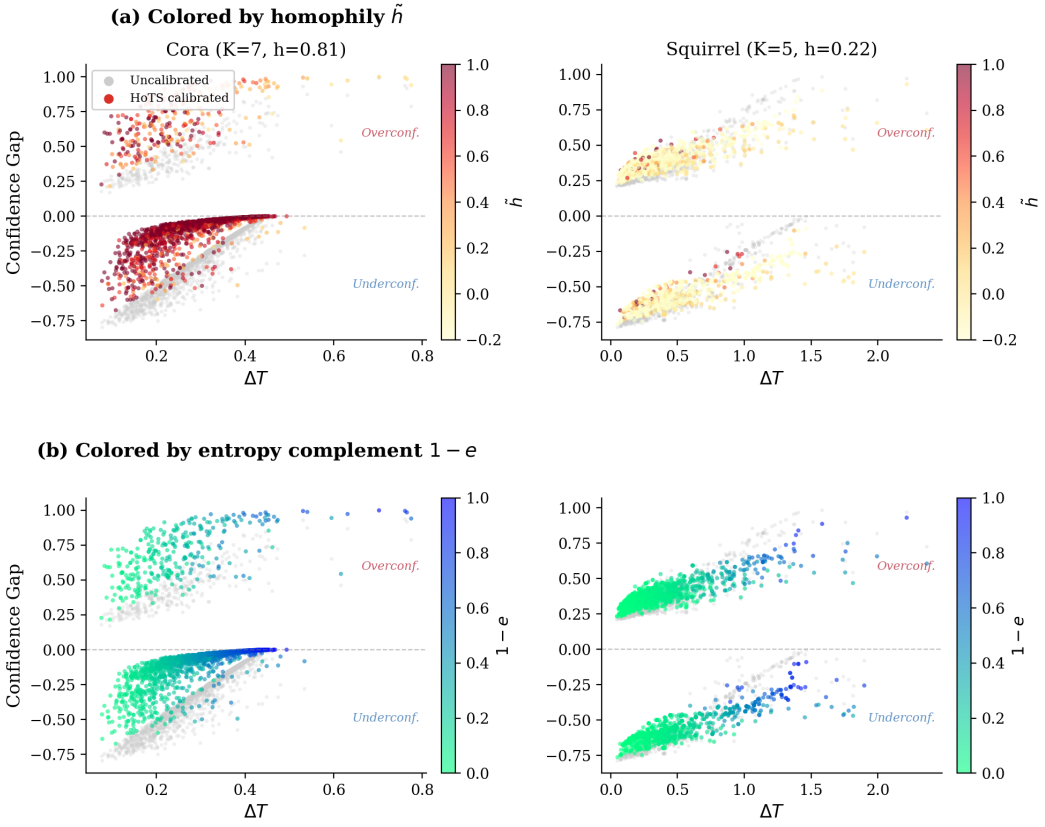

Figure 4: Node-level diagnostic for HoTS on representative datasets. Points show confidence gap versus the homophily-dependent temperature correction evaluated with oracle homophily. The two rows color the same nodes by oracle normalized homophily $\tilde{h}$ and entropy complement $1 - e$. Oracle homophily is computed from true labels solely for this diagnostic.

Figures 5 and 6 provide the same diagnostic across all datasets.

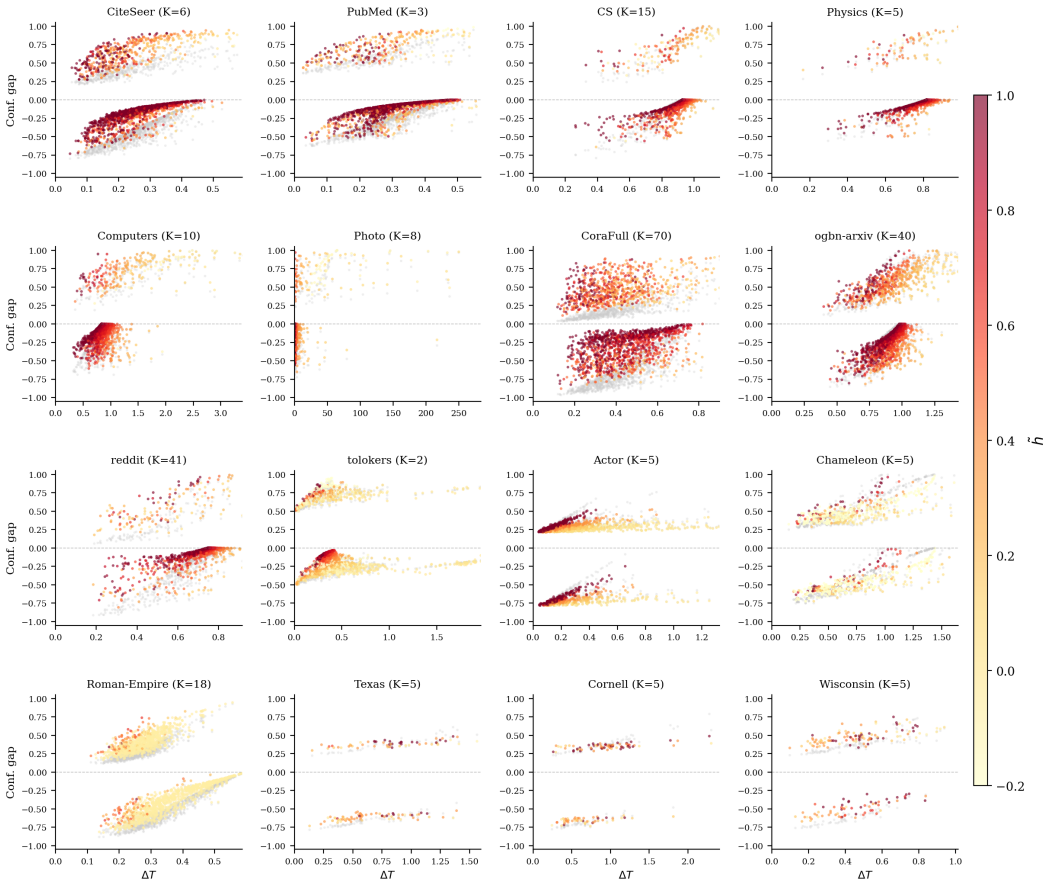

Figure 5: Node-level diagnostic across datasets, colored by oracle normalized homophily $\tilde{h}$, computed from true labels for this diagnostic. The effect of HoTS is most structured on datasets where local homophily varies substantially across nodes.

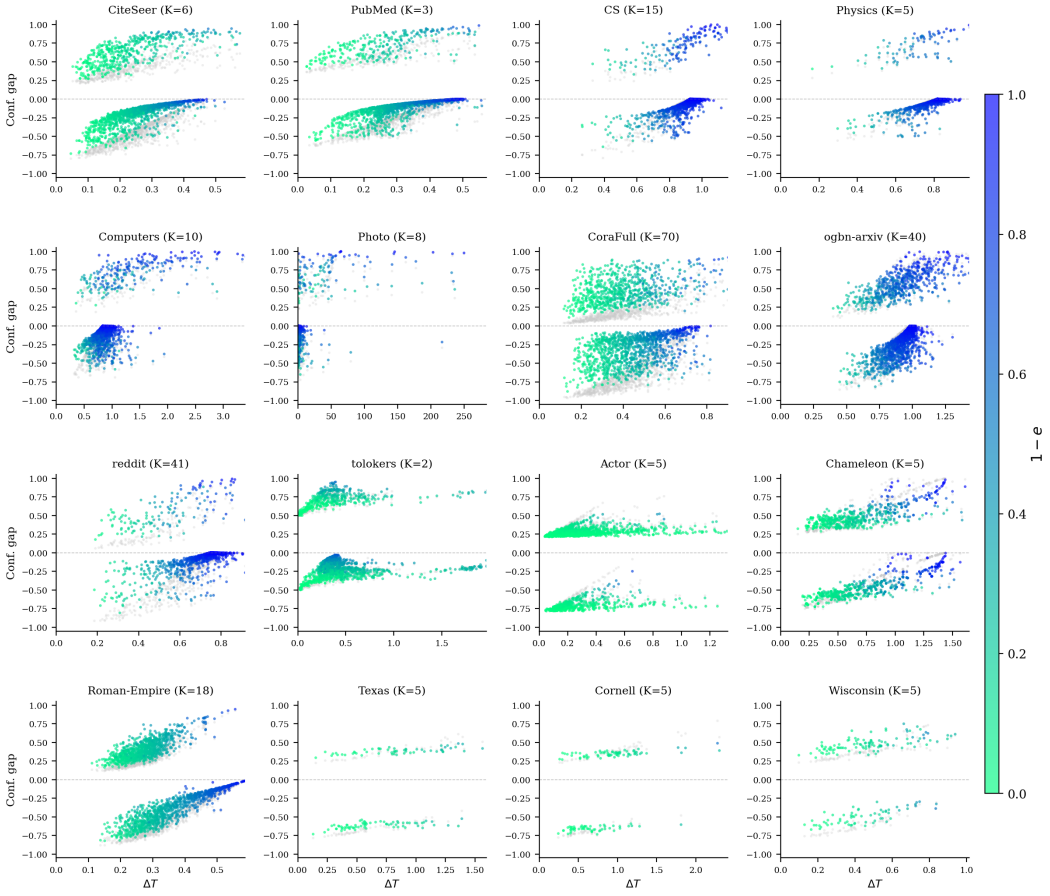

Figure 6: Node-level diagnostic across datasets, colored by entropy complement $1 - e$. Entropy complement captures logit concentration and is complementary to the homophily signal in Figure 5.

B.5 KL Bound Validation

Appendix A.8 derives that, in the sparse-overlap regime, the Gaussian residual KL from the joint neighbor-logit law to the product law scales as $O(K^2\bar{d}^2/n)$. Figure 7 checks this scaling on CSBM graphs using three independent sweeps. The fitted slopes are close to the predicted exponents. The sweep over K has slope 1.81 overall and 1.97 when restricted to $K \geq 8$, the sweep over $\bar{d}$ has slope 2.08, and the sweep over n has slope -1.01. A discrimination test at fixed n_{pc} favors quadratic rather than linear scaling in K.

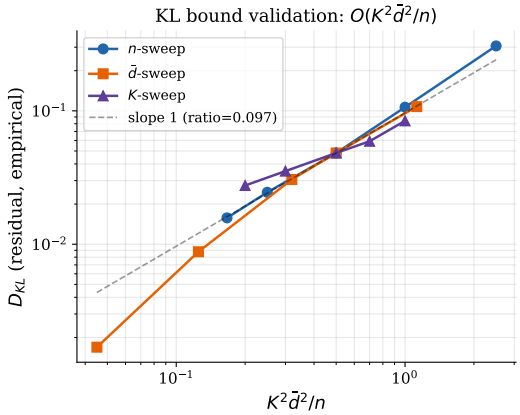

Figure 7: KL bound validation on CSBM. Three independent sweeps over n, $\bar{d}$, and K collapse when plotted against $K^2\bar{d}^2/n$, supporting the residual scaling derived in Appendix A.8.

B.6 Structural Perturbation Stress Test

We also evaluate a simple structural stress test. Edges are randomly added or removed at rates from 5% to 50%. This auxiliary experiment uses predicted-label neighbor agreement to estimate homophily, rather than the learned GCN estimator used in the main experiments. Logits and this estimate are recomputed on each perturbed graph, and TS and HoTS are refitted on its calibration split. This is not a full cross-domain distribution shift study, but it tests whether the structural feature used by HoTS becomes brittle under moderate graph perturbations. Table 8 reports ECE on representative homophilic and heterophilic benchmarks under increasing perturbation rates.

Table 8: ECE (%) under edge perturbation. HoTS is comparatively stable on the homophilic benchmarks shown here. On heterophilic graphs, random perturbation can improve calibration for both TS and HoTS by weakening the original heterophilic structure.

Dataset	Method	0%	5%	10%	20%	30%	50%
Cora	HoTS	2.27	2.72	2.21	2.55	2.77	3.03
	TS	3.39	3.24	2.81	3.73	3.69	4.16
CiteSeer	HoTS	2.83	3.09	3.21	3.10	2.72	3.92
	TS	3.09	2.95	3.09	3.11	3.13	3.79
CS	HoTS	0.99	0.95	0.95	1.26	1.19	1.15
	TS	1.49	1.49	1.54	1.93	1.89	2.44
Chameleon	HoTS	9.58	6.50	7.28	6.33	5.90	5.36
	TS	12.61	10.00	10.22	7.45	5.90	3.29
Squirrel	HoTS	5.09	3.27	2.46	1.68	1.25	0.66
	TS	7.85	7.30	4.71	3.02	2.10	3.93

On Cora and CS, HoTS changes by less than one percentage point at 50% perturbation, whereas TS degrades more substantially. On Chameleon and Squirrel, calibration improves for both methods as random perturbations partially wash out the original heterophilic structure. This mixed behavior is consistent with the role of α. When estimated homophily is informative, HoTS can exploit it. When

the structural signal is weakened or altered, the learned exponent and the base temperature reduce the reliance on that signal.

B.7 Confidence-Conditioned Homophily Diagnostics

We use oracle local homophily to examine whether neighborhood label composition is associated with prediction correctness within base-model confidence bins. This choice isolates the structural quantity motivating Proposition 1 from errors introduced by the learned homophily estimator. Oracle homophily h_i is the fraction of same-label neighbors on the backbone input graph, including self-loops. We pool GCN test predictions over ten seeds per dataset, stratify nodes into five confidence-quantile bins, and compare empirical accuracy across oracle-homophily groups within each bin. Figure 8 uses a 50/50 split, while Figure 9 uses three groups. True labels are used solely for this retrospective grouping; these oracle values are not used to fit the homophily estimator or the temperature map. The implemented HoTS uses the observed-label-only estimate $\hat{h}_i$.

B.8 Comparison with Other Structural Statistics

To examine the empirical basis for focusing on homophily, we compare its residual association with prediction correctness against representative motif, centrality, and curvature statistics. We use oracle normalized local homophily, the local clustering coefficient, PageRank, and mean incident augmented Forman–Ricci curvature. Oracle homophily is computed from true labels solely for this retrospective diagnostic; it is not an input to the HoTS predictor or temperature map.

We evaluate GCN predictions over five seeds. For each dataset and seed, we separately regress the correctness indicator and each structural statistic on an intercept, uncalibrated confidence, and $\log(1+d_i)$, where d_i is the in-degree in the backbone input graph. Their residual correlation measures the additional linear association with correctness after controlling for confidence and degree. The pooled summary concatenates residuals across datasets and seeds; the macro summary averages dataset-level correlations, each computed after pooling that dataset's residuals across seeds.

Clustering, PageRank, and curvature are computed on the undirected, self-loop-free graph. PageRank uses damping 0.85. For an edge (u, v), augmented Forman–Ricci curvature is $4 - d_u - d_v + 3t_{uv}$, where degrees and the common-neighbor count t_{uv} refer to this graph; node curvature averages over incident edges. Oracle homophily is the same-label outgoing-neighbor fraction on the backbone input graph, including self-loops, normalized as $\tilde{h}_i = (Kh_i - 1)/(K - 1)$. Homophily and PageRank cover all benchmarks; clustering and curvature omit Reddit because of the cost of triangle computations.

Table 9: Partial correlation with prediction correctness after controlling for confidence and log-degree. Results use GCN and five seeds. Homophily and PageRank cover 18 datasets; clustering and curvature cover 17, excluding Reddit.

Structural statistic	Category	Pooled	Macro
Oracle normalized homophily	Homophily	$+0.384$	$+0.283$
Local clustering coefficient	Motifs	$+0.044$	$+0.012$
PageRank	Centrality	$+0.003$	-0.003
Mean incident augmented Forman–Ricci curvature	Curvature	-0.026	-0.052

Among the statistics examined, oracle homophily shows the strongest residual association with correctness in both summaries (Table 9). Together with the explicit homophily-dependent temperature law from the CSBM analysis, this provides an empirical basis for our choice of structural variable. The comparison concerns the structural properties themselves and their conditional linear association with correctness.

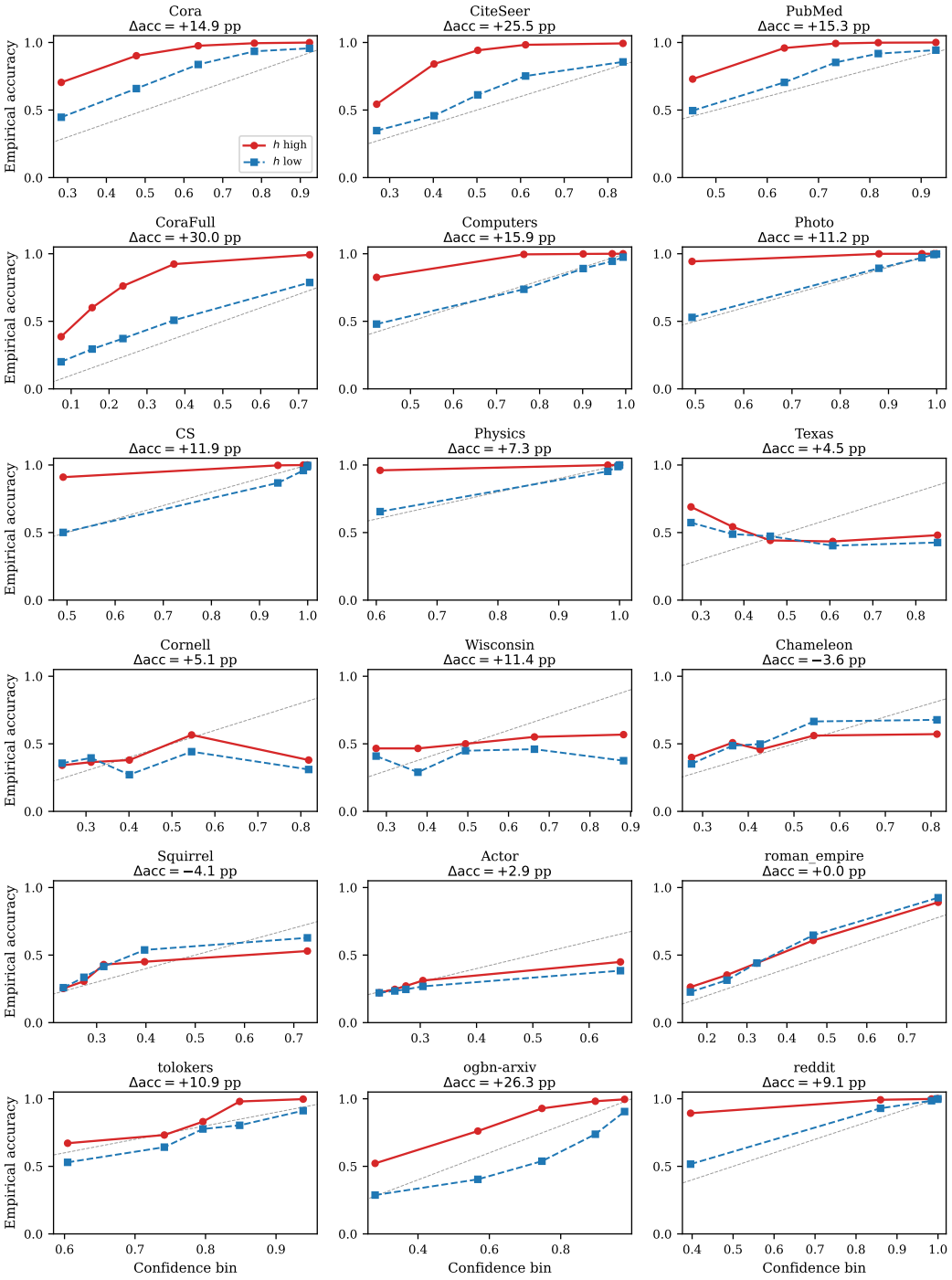

Figure 8: Confidence-conditioned oracle-homophily diagnostic with a 50/50 split. For each dataset, test nodes are stratified by the base-model confidence and separated into low- and high-h groups using a rank-based 50/50 split of oracle local homophily within each bin; ties retain the pooled node order. Oracle homophily is used to examine the structural reliability signal separately from estimation error. Curves show empirical accuracy within each confidence bin. The title of each panel reports the mean high-minus-low empirical accuracy gap Δ_{acc} across the five confidence bins in percentage points.

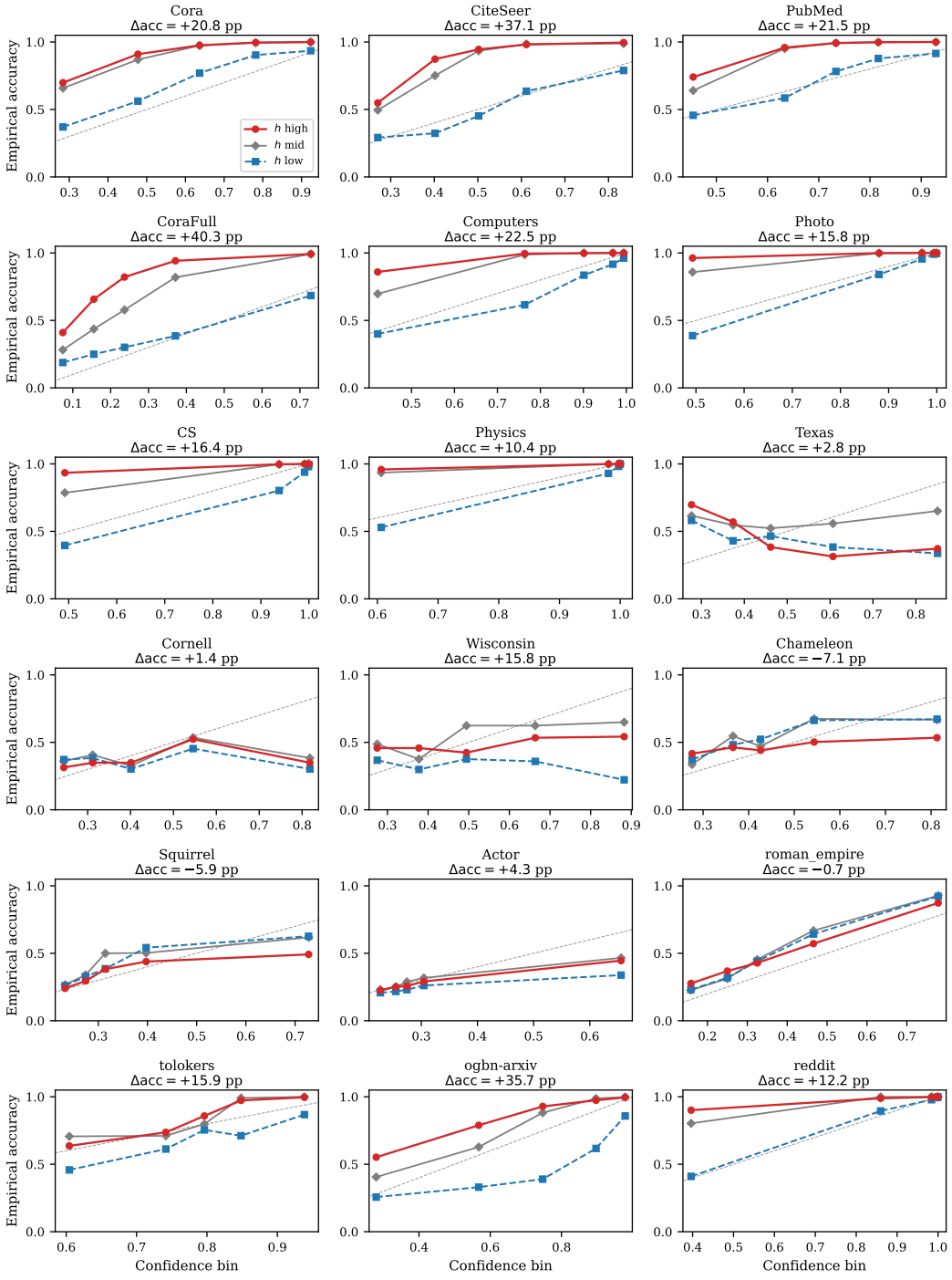

Figure 9: Confidence-conditioned oracle-homophily diagnostic with three groups. This figure repeats the diagnostic in Figure 8, but separates nodes by oracle-homophily rank into low-, middle-, and high-h groups of approximately equal size within each confidence bin. The title of each panel reports the mean high-minus-low empirical accuracy gap Δ_{acc} across the five confidence bins in percentage points. The three-way split visualizes how accuracy varies with neighborhood label composition at comparable confidence.

B.9 HoTS Component Ablation

Component ablation. Table 10 compares TS, entropy-only scaling, homophily-only scaling, HoTS with a fixed exponent, and the full HoTS. TS uses a global temperature. Entropy-only uses $T_i = T_{\text{base}} + \beta\sqrt{2K \log K(1 - e_i)}$, omitting the homophily denominator and fitting its own positive scalar parameters. Homophily-only uses $T_i = T_{\text{base}} + \beta/(|\hat{\tilde{h}}_i| + \varepsilon)^{\alpha}$, omitting the entropy numerator. HoTS combines both terms as defined in Section 4, while HoTS ($\alpha = 1$) retains this combination but fixes the exponent. HoTS achieves the lowest reported mean ECE.

Table 10: HoTS component ablation. Mean ECE (%) across 18 datasets, pooling GCN and GAT with 10 seeds per backbone. HoTS and HoTS ($\alpha = 1$) use the same observed-label-only homophily estimates. †Homophily-only retains the previously reported result and has not been rerun with the current observed-label-only estimator. Lower is better; best in **bold**.

	TS	Entropy-only	Homophily-only†	HoTS ($\alpha = 1$)	HoTS (ours)
Mean ECE (%)	6.04	4.83	5.56	5.90	**4.79**

Learned versus fixed exponent. The fixed-$\alpha = 1$ variant refits T_{base} and β using the same backbone checkpoints, homophily estimates, splits, and optimization settings as HoTS. Its higher mean ECE supports learning the homophily exponent rather than using the oracle asymptotic value on noisy real graphs.

C Implementation Details

This appendix collects the experimental and implementation details needed to reproduce our results, including dataset statistics and preprocessing, hardware and software environment, training hyperparameters for the GNN backbones and the calibration methods, the architecture of the HoTS homophily predictor, and a measurement of prediction preservation across calibrators.

C.1 Dataset Statistics and Preprocessing

Table 11 reports the size, feature dimension, number of classes, and node-level homophily for the 18 benchmark datasets used in our experiments. Homophily is computed as the mean ($\pm$ std) of node-level homophily $h_i = d_i^{-1} \sum_{j \in \mathcal{N}(i)} \mathbf{1}\{y_j = y_i\}$ across all nodes (with $h_i = 0$ for isolated nodes).

Table 11: Dataset statistics. Homophily is the mean ($\pm$ std) of node-level homophily h_i over all nodes. All datasets use random 20/10/70 train/val/test splits regenerated for each seed.

Dataset	Nodes	Edges	Classes	Features	Homophily
Citation (Planetoid)					
Cora	2,708	13,264	7	1,433	0.871 ± 0.211
CiteSeer	3,327	12,431	6	3,703	0.820 ± 0.245
PubMed	19,717	108,365	3	500	0.865 ± 0.217
Amazon / Coauthor					
Computers	13,381	491,556	10	767	0.802 ± 0.239
Photo	7,487	238,087	8	745	0.849 ± 0.227
CS	18,333	163,788	15	6,805	0.832 ± 0.239
Physics	34,493	495,924	5	8,415	0.915 ± 0.174
CoraFull	18,800	144,170	70	8,710	0.677 ± 0.274
WebKB					
Texas	183	492	5	1,703	0.530 ± 0.268
Cornell	183	478	5	1,703	0.560 ± 0.273
Wisconsin	251	750	5	1,703	0.565 ± 0.290
Wikipedia					
Chameleon	2,277	36,101	5	2,325	0.104 ± 0.229
Squirrel	5,201	217,073	5	2,089	0.089 ± 0.194
Actor	7,600	40,869	5	932	0.608 ± 0.319
Heterophilous Benchmark					
Roman-Empire	22,662	65,854	18	300	0.046 ± 0.139
tolokers	11,758	1,038,000	2	10	0.634 ± 0.247
Large-scale					
ogbn-arxiv	169,343	2,501,829	40	128	0.707 ± 0.272
Reddit	232,965	114,848,857	41	602	0.813 ± 0.230

Preprocessing. We apply three preprocessing steps before training: **(i) Largest connected component (LCC).** For Amazon (Computers, Photo) and Coauthor / Coauthor-Full (CS, Physics, CoraFull) we extract the largest connected component, following Zhuang et al. [2025], so that all nodes are reachable during message passing. **(ii) Undirected edges.** Directed graphs (in particular ogbn-arxiv) are converted to undirected by adding reverse edges. **(iii) Self-loops.** A self-loop is added at every node, which is standard for GCN-style aggregation [Kipf and Welling, 2017] and ensures that a node's own features participate in the layer-wise update.

Splits and protocol uniformity. For each (dataset, seed) we sample a fresh random 20/10/70 train/validation/test split. The validation split is used for both early stopping of the GNN backbone and for fitting all post-hoc calibrators. The test split is used only for the reported metrics. We use random splits across all 18 datasets to keep the calibration evaluation protocol consistent across heterogeneous

benchmarks; this also matches the protocol assumed in the public reference implementations of the graph-aware baselines (CaGCN, GATS, GETS, WATS). GATS is omitted on Reddit because it exceeds the 24 GB GPU memory budget under this protocol (Appendix C.2).

C.2 Hardware and Software

All experiments are run on a workstation with 4× NVIDIA RTX A5000 GPUs (24 GB each); each individual run uses a single GPU. The software stack is Python 3.10, PyTorch 2.4.0 with CUDA 12.1, DGL 2.4.0, PyTorch Geometric 2.7.0, and OGB 1.3.6. Random seeds for Python, NumPy, PyTorch, and DGL are fixed before each (dataset, backbone, seed) run, and we set `CUBLAS_WORKSPACE_CONFIG=:4096:8` together with `torch.use_deterministic_algorithms` for reproducible CUDA kernels.

Memory limits. GATS results on Reddit are reported as "–" because its attention-based per-node temperature module exceeds the 24 GB GPU memory budget on this graph (Tables 1 and 4).

C.3 GNN Training Hyperparameters

The main experiments use two backbones: a 2-layer GCN [Kipf and Welling, 2017] and a 2-layer GAT [Veličković et al., 2017]. Both backbones are trained from scratch for each (dataset, seed) pair using the cross-entropy loss on the training split, with early stopping on the validation loss. The GAT uses 2 attention heads with attention and feature dropout both equal to the GNN dropout. Hyperparameters follow the per-dataset settings adopted by prior calibration work [Zhuang et al., 2025] and group into three regimes summarized in Table 12.

Dataset groups. (G1) Cora, CiteSeer, PubMed, Texas, Cornell, Wisconsin, Chameleon, Squirrel, Actor, Roman-Empire, tolokers, and Reddit. (G2) Computers, Photo, CS, Physics, CoraFull. (G3) ogbn-arxiv.

Table 12: GNN training hyperparameters by dataset group. Optimizer is Adam in all cases; both backbones train for at most 200 epochs; "–" under Patience indicates no early stopping (full 200 epochs are run). The same configuration is used for GCN and GAT, with GAT additionally using 2 attention heads and feature/attention dropout equal to the listed Dropout. Group G3 (ogbn-arxiv) additionally uses BatchNorm between layers, following the OGB convention.

Group	Hidden	Layers	LR	Weight decay	Dropout	Patience
G1	16	2	10^{-2}	5×10^{-4}	0.5	50 / –
G2	64	2	10^{-2}	10^{-3}	0.5	50
G3	256	2	10^{-2}	0	0.5	–

Common settings. All runs use the Adam optimizer, train for at most 200 epochs, and use the validation split both for early stopping and for fitting all post-hoc calibrators. ogbn-arxiv additionally uses BatchNorm between GCN/GAT layers, following the OGB leaderboard convention. PubMed, ogbn-arxiv, and Reddit run all 200 epochs without early-stopping patience; the remaining datasets stop when the validation loss has not improved for 50 epochs.

C.4 Calibration Method Hyperparameters

With the GNN backbone parameters frozen, the iterative calibration stage optimizes parameters on the validation (calibration) split. We use the Adam optimizer with learning rate 10^{-2} and no weight decay, train for at most $1{,}000$ epochs, and use training-split cross-entropy for checkpoint selection and early stopping with a patience of 50 epochs. Best weights from the early-stopping checkpoint are restored before evaluation. The separate HoTS homophily predictor follows the supervision and training protocol in Appendix C.5. These common iterative calibration settings apply to all methods listed below; method-specific architectural details are summarized in Table 13.

Table 13: Method-specific calibration settings. Common settings (Adam, lr=10^{-2}, weight decay 0, 1,000 epochs, patience 50) apply to all methods. "Per-dataset config" indicates that learning rate, hidden width, attention heads, or other architectural choices follow the per-dataset configuration files released with the original method.

Method	Method-specific settings
TS, ETS, HTS	Scalar parameters only; no architectural hyperparameters.
VS	Per-class affine logit transform $z_c \mapsto w_c\, z_c + b_c$.
CaGCN, GATS	Per-dataset config (auxiliary GCN/attention width and dropout) following the original releases.
GETS	Per-dataset config (mixture-of-experts width and number of experts) following the original release.
WATS	Per-dataset config (graph wavelet basis size) following the original release.
HoTS (ours)	Three scalar temperature parameters $(T_{\text{base}}, \beta, \alpha)$, enforced through softplus parameterizations $T_{\text{base}} = \text{softplus}(\hat{T}) + 0.1$, $\beta = \text{softplus}(\hat{\beta}) + 0.01$, $\alpha = \text{softplus}(\hat{\alpha}) + 0.01$ on unconstrained underlying parameters $\hat{T}, \hat{\beta}, \hat{\alpha} \in \mathbb{R}$. We initialize $\hat{T} = \hat{\beta} = 0.5$ and $\hat{\alpha} = 0$, giving effective initial values $T_{\text{base}} \approx 1.07$, $\beta \approx 0.98$, $\alpha \approx 0.70$. The stabilizer in the homophily denominator is fixed to $\varepsilon = 0.02$. The homophily predictor that produces $\hat{h}_i$ is trained separately; see Appendix C.5.

C.5 HoTS Homophily Predictor

The HoTS denominator depends on the normalized estimate $\hat{\tilde{h}}_i = (K\hat{h}_i - 1)/(K-1)$. We train an auxiliary GCN on homophily targets constructed solely from observed training and validation labels. The predictor is fitted once within each (dataset, backbone, seed) calibration run, and its predictions are cached before fitting the temperature map. The base classifier and its logits remain fixed throughout calibration.

Architecture. The predictor is a 2-layer GCN of the form

$$\hat{h}_i = \sigma(\text{GCN}_2(\text{ReLU}(\text{GCN}_1(x_i; A))\,;\, A))\,,$$

mapping from input feature dimension F to a hidden width of 32 and then to a scalar output passed through a sigmoid so that $\hat{h}_i \in (0, 1)$. The same architecture is used for all 18 datasets.

Supervision target. Let $\mathcal{I}_{\text{known}} = \mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{val}}$ be the nodes with observed labels. Define the observed-label neighborhood and the supervised node set as

$$\mathcal{N}_L(i) = \{j \in \mathcal{N}(i) : j \in \mathcal{I}_{\text{known}},\ j \neq i\},$$
$$\mathcal{M} = \{i \in \mathcal{I}_{\text{known}} : |\mathcal{N}_L(i)| > 0\}.$$

For $i \in \mathcal{M}$, the masked supervision target is

$$h_i^{\text{mask}} = \frac{1}{|\mathcal{N}_L(i)|} \sum_{j \in \mathcal{N}_L(i)} \mathbf{1}\{y_j = y_i\}.$$

The target uses only observed-label neighbors, excluding self-loops. Known nodes without such neighbors are excluded from the predictor loss.

Predictor training. The auxiliary GCN minimizes

$$\mathcal{L}_{\text{pred}}(\theta) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(\hat{h}_i(\theta) - h_i^{\text{mask}}\right)^2$$

using Adam with learning rate 10^{-2} and weight decay 10^{-4} for at most 200 epochs. Early stopping monitors this supervised training loss with a patience of 30 epochs, and the best-loss checkpoint is restored. The GCN uses the full transductive graph with self-loops and node features to predict homophily for all nodes, including those without observed labeled neighbors.

Temperature fitting and label access. After predictor training, we compute and cache $\hat{h}_i$ for every node and normalize it to $\hat{\tilde{h}}_i = (K\hat{h}_i - 1)/(K-1)$. These estimates are held fixed while fitting $T_{\text{base}}, \beta, \alpha$ on the frozen classifier logits. The temperature map minimizes validation-split cross-entropy; training-split cross-entropy selects the checkpoint and controls early stopping. Test labels are used only for evaluation and retrospective diagnostics, not for fitting or checkpoint selection.

C.6 Prediction Preservation Across Calibrators

A node-wise calibrator that preserves the predicted class is desirable: any improvement in selective classification or downstream decisions should come from better confidence *ranking*, not from changing the underlying classifier. To quantify this, we measure, for each calibrator and for each of the $360 = 18\times2\times10$ (dataset, backbone, seed) experiments, whether at least one test node's argmax differs from the uncalibrated baseline. Table 14 reports the resulting rate. Methods that apply a positive scalar temperature to logits ($T_i > 0$ for every node) are guaranteed to preserve $\arg\max$; methods that learn class-specific or per-class affine transforms (VS, GETS) can reorder logit components and change the predicted class.

Table 14: Prediction change rates across all 360 experiments (18 datasets $\times$ 2 backbones $\times$ 10 seeds). "Preserves" indicates whether the method theoretically preserves the predicted class. "Rate" is the fraction of experiments in which at least one test node's $\arg\max$ differs from the uncalibrated baseline. Lower is more conservative. *ETS can in principle reweight class probabilities but empirically learns mixture weights very close to $(1, 0, 0)$. †CaGCN and GATS produce strictly positive T_i, but rare numerical edge cases (near-tied logits) flip a small number of predictions. ‡Reddit GATS is omitted due to OOM during fitting (Appendix C.2); the rate for GATS is computed over the remaining 340 experiments.

Method	Calibrator form	Preserves	Rate
Global methods			
TS	$\mathrm{softmax}(z/T)$, $T > 0$ scalar	Yes	0.0%
ETS	$\sum_k w_k\, p^{(k)}$, mixture	No*	0.3%
VS	$z_c^{\text{cal}} = w_c z_c + b_c$ per-class affine	No	96.1%
Node-wise methods			
HTS	$\mathrm{softmax}(z_i/T_i)$, $T_i = a + b\, e_i$, $T_i > 0$	Yes	0.0%
CaGCN	$\mathrm{softmax}(z_i/T_i)$, $T_i = \mathrm{GCN}(x, A)_i$	Yes†	1.9%
GATS	$\mathrm{softmax}(z_i/T_i)$, $T_i = \mathrm{Attn}(x, A)_i$	Yes†	0.9%‡
GETS	$z_{i,c}^{\text{cal}} = z_{i,c} \cdot \mathrm{softplus}(T_{i,c})$	No	98.6%
WATS	$\mathrm{softmax}(z_i/T_i)$ with wavelet-based $T_i > 0$	Yes	0.0%
HoTS	$\mathrm{softmax}(z_i/T_i)$, $T_i = T_{\text{base}} + \frac{\beta\sqrt{2K\log K(1-e_i)}}{(\lvert\hat{\tilde{h}}_i\rvert + \varepsilon)^\alpha} > 0$	Yes	**0.0%**

HoTS, like TS, HTS, and WATS, never changes the predicted class because $T_i > 0$ is enforced by the softplus parameterization (Appendix C.4). This means the gains in Table 1, Table 3, and Table 2 reflect improved probability ranking rather than reclassification. In contrast, VS and GETS modify the predicted class on roughly 96% and 99% of experiments respectively, so any selective-classification advantage they may show would be confounded with prediction refinement and we therefore exclude them from Table 2.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The four contributions stated at the end of Section 1 correspond to Proposition 1 (Section 3), Theorem 1 (Section 3, proof in Appendix A.3), Section 4, and Section 5 respectively. The aggregate empirical claims (best mean ECE 4.79%, best average rank, and the most reliable confidence ranking in selective classification) match Tables 1 and 2; their scope is qualified in Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the theoretical scope (CSBM/Gaussian/equidistant-means/uniform-prior/one-layer-linear-GCN/population-concentration assumptions) and challenges from class imbalance, nonlinear multi-layer message passing, and low-degree finite-sample fluctuations. The positive-signal restriction with the absolute-value extension is detailed in Section 3 and Appendix A.3, and the surrogate reliability-matching step is described in Section 3. The GATS out-of-memory failure on Reddit is disclosed in Section 5 and Appendix C.2.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are stated in Section 3 and formalized in Appendix A.1. Proposition 1 is proved in Appendix A.2, Theorem 1 in Appendix A.3, and Proposition 2 in Appendix A.7; the scope of each result (positive-signal region $\mathcal{H}_+$, large-$\widetilde{dh}$ regime) is stated alongside the corresponding theorem.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 states the high-level setup; Appendix C provides full implementation details (datasets in C.1, hardware/software in C.2, GNN training in C.3, calibration hyperparameters in C.4, homophily predictor in C.5). The HoTS formula is given in Equation (8).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at `https://github.com/inu0104/HoTS`. The repository contains the `hots` package (data loaders, GNN backbones, and calibrators), dataset and calibrator configuration files, and `run_experiment.py`, which handles backbone training, calibration, and evaluation; the public benchmarks are downloaded automatically. `README.md` gives the installation steps and the exact commands for a single experiment and for the main benchmark (18 datasets, GCN and GAT, seeds 0–9), and the dependencies are pinned in `pyproject.toml` (Python 3.10, PyTorch 2.4.0, DGL 2.4.0, CUDA 12.1).

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 5 states data splits, seeds, backbones, metrics, and baselines; Appendices C.3, C.4, and C.5 give per-group GNN training hyperparameters, calibration optimizer settings, and the homophily predictor details. Hyperparameters follow the per-dataset settings of Zhuang et al. [2025].

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All per-dataset cells in Tables 1, 4, 5, and 6 report mean $\pm$ sample standard deviation over 10 random seeds per backbone (the pooled GCN+GAT main table aggregates over both backbones); variability is over seeds (each seed regenerates the split, reinitializes the GNN, and refits the calibrator).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.2 reports the hardware ($4\times$ NVIDIA RTX A5000 24 GB; one GPU per run) and the software stack (PyTorch 2.4.0, CUDA 12.1, DGL 2.4.0, PyTorch Geometric 2.7.0, OGB 1.3.6). The GATS calibrator on Reddit exceeds 24 GB of GPU memory, which is disclosed.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. The work is post-hoc calibration on public node-classification benchmarks; it does not involve human subjects, sensitive personal data, or generative content.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Better-calibrated GNN confidence supports more reliable rejection thresholds and human-in-the-loop review in deployed graph learning systems (positive impact). We do not expect direct negative impacts from a method that improves probability reliability while preserving the underlying classifier's predictions; however, well-calibrated scores should not be interpreted as a license for greater automation in high-stakes decisions, since post-hoc calibration does not by itself improve the underlying model's accuracy or fairness across protected groups, which should be evaluated separately before deployment.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release pretrained generative models, scraped data, or other high-misuse-risk artifacts; the contribution is a post-hoc calibrator with no such risks.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets are public node-classification benchmarks accessed through DGL, PyTorch Geometric, or OGB, and are cited at first use in Section 5 with their original sources [Yang et al., 2016, Shchur et al., 2018, Pei et al., 2020, Platonov et al., 2023, Hu et al., 2020, Hamilton et al., 2017]. Baseline calibrators are reused from the public reference implementations of the corresponding papers [Wang et al., 2021, Hsu et al., 2022, Zhuang et al., 2025, Li et al., 2025]; we use them under their stated license terms.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset is the HoTS calibrator and its experiment pipeline, released at `https://github.com/inu0104/HoTS`. Documentation includes `README.md` (installation, automatic dataset download, exact reproduction commands), `pyproject.toml` (pinned dependencies), dataset and calibrator configuration files, and an MIT `LICENSE` file. No new dataset is introduced; all 18 benchmarks are public.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects; all evaluation uses public node-classification benchmarks.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects, so IRB approval is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not a component of the core methodology; HoTS uses three scalar parameters and a small auxiliary GCN-based homophily predictor, and no analysis or experiment depends on any LLM.