Learning Curves of GNNs vs MLP vs Tikhonov

A Comparative Study on Semi-Supervised Node Classification

Calin Radoi

Responsible Professor: Elvin Isufi
Supervisors: Chengen Liu, Mohamed Jebali

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Abstract

Graph neural networks (GNNs) are designed to use both node attributes and graph topology, but this extra information is not always beneficial. This paper compares learning curves for semi-supervised node classification under five informational and structural regimes: a ChebNet GNN (implicit topology and features), a feature-only multilayer perceptron (MLP), a topology-only graph Tikhonov baseline, and two variations enforcing explicit topological prior information via Graph Laplacian Regularisation (a regularized MLP and a regularized GNN). We evaluate Cora and PubMed as homophilic citation networks, and Chameleon and Squirrel as heterophilic web graphs over increasing labeled-node budgets, using fixed validation and test splits and repeated random budget sampling. We find that in the homophilic settings, the structural benefit of the GNN is substantial, though enforcing a strict smoothness prior makes the Tikhonov baseline surprisingly competitive at extremely low label budgets. In heterophilic settings, where the underlying structural smoothness prior is mathematically violated, model behaviour bifurcates based on structural density. On sparser heterophilic graphs (Chameleon), all informational regimes converge to an identical performance floor bounded by uninformative node features. Conversely, on massively dense heterophilic networks (Squirrel), classical topology-only solvers catastrophically collapse below feature-only baselines due to the amplified friction of the deceptive prior. Our results show that the benefit of explicit topological priors depends heavily on data availability. They provide clear advantages in low-label regimes under strict feature starvation, but become less effective as more labels or stronger feature signals become available.

1 Introduction

Graphs are a natural representation for citation networks, web pages, social systems, and other domains in which examples are not independent [6, 13]. Graph neural networks (GNNs) exploit this structure by combining node features with information propagated along edges, and have become a standard tool for semi-supervised node classification [4, 9]. Yet a higher-capacity model is not automatically a better model: when labels are scarce, performance may depend more on the sampling of labelled nodes and on the relation between graph topology and class labels than on architecture alone.

This distinction matters because graph structure can be either helpful or misleading. In homophilic graphs, neighbours often share labels, so smoothing predictions across edges is a useful inductive bias. In heterophilic graphs, neighbours often belong to different classes, so the same bias can wash out discriminative information [11, 17]. To systematically capture these dynamics, we utilize learning curves [14], which evaluate model performance as a continuous function of the number of labelled training nodes (n_l). A learning curve is therefore more informative than a single benchmark number: it reveals data efficiency, showing whether graph structure helps only in low-label regimes, whether feature-only models eventually catch up, and whether topology-only smoothing remains competitive as more labels become available.

Despite the proliferation of specialized GNN architectures designed to handle varying structural regimes, the current literature predominantly evaluates model performance at a single, static label rate [12]. This heavy reliance on single-point benchmarks creates a critical scientific gap: the field lacks a continuous understanding of data efficiency. It remains unclear whether topological message passing provides a persistent advantage across all stages of data availability, or if feature-only models eventually catch up as the label budget increases. By adopting the continuous learning curve framework, this thesis directly addresses this gap, providing a dynamic evaluation of structural utility that static benchmarks obscure.

This paper asks: *when do graph structure and node features synergise implicitly, when are explicit topological priors necessary, and when do these mechanisms become redundant or actively detrimental?* We address this research question by systematically evaluating learning curves under five distinct informational regimes: a standard spectral ChebNet [4] (implicit topology and features), a feature-only Multilayer Perceptron (MLP), a topology-only Graph Tikhonov baseline, and two variations enforcing explicit topological prior information via Graph Laplacian Regularisation (a regularised MLP and a regularised GNN).

By tracing performance across both homophilic (Cora, PubMed [13]) and heterophilic (Chameleon, Squirrel [11]) graphs at multiple label budgets, we isolate the structural benefit gap and provide a nuanced view of where and why graph structure improves classification. We further anchor our empirical findings in classical Graph Signal Processing theory [8], quantifying dataset signal smoothness via Dirichlet Energy to explain the mathematical success or universal convergence to a baseline performance floor of explicit topological priors, and tracking optimization gradients to investigate architectural regularisation stability.

The remainder of this paper is structured as follows. Section 2 positions our work within the existing literature of spectral graph architectures, manifold regularisation, and graph homophily. Section 3 formalizes the semi-supervised node classification problem, outlines our five-model evaluation framework, and mathematically defines our analytical tracking metrics. Section 4 details our experimental setup, dataset structural properties, and hyperparameter choices. Section 5 presents our primary empirical learning curves alongside an integrated discussion of their theoretical implications, analyzing the phenomena of feature-dominance, catastrophic mixing, and algorithmic redundancy. Section 6 details our responsible research practices, and Section 7 concludes the paper and maps out strategic directions for future work. Finally, Appendix A outlines our baseline hyperparameter optimization grid search, and Appendix B provides extended training trajectories via gradient cosine similarity diagnostics for further research tracking.

2 Background and Related Work

Implicit Forward Aggregation vs. Explicit Regularisation. Spectral graph neural networks define convolution through graph operators, making the graph Laplacian central to the model class [2]. ChebNet made this idea practical by approximating spectral filters with Chebyshev polynomials, localising the filter without an explicit eigendecomposition [4]. Standard GNN benchmarking typically focuses on comparing different network architectures [9]. However, few studies separate the effects of sharing graph structure implicitly during the forward pass (via feature mixing) versus enforcing structural penalties explicitly during training (via loss priors). Classical graph signal processing (GSP) enforces structural constraints explicitly; manifold regularisation and label smoothing [1, 16] operate as prior-based models that directly penalise outputs violating a homophilic smoothness assumption. It remains unclear whether adding an explicit backward-pass penalty provides any extra benefit when a GNN's layers already mix neighbouring information automatically.

The Boundaries of Homophily and Graph Density. Standard GNN benchmarking often relies on citation networks (e.g., Cora, PubMed [13]) where the homophily assumption strictly holds. While recent literature explores heterophilic web graphs where this assumption breaks down [11, 17], evaluating these datasets purely on static performance obscures the underlying structural mechanics. The failure of topological solvers on heterophilic graphs is heavily exacerbated by graph density, which induces catastrophic mixing when conflicting neighbourhood signals are analytically smoothed. By applying classical GSP smoothness priors to filtered, densely connected heterophilic networks [12], we can isolate how neural optimisation landscapes react to mathematically toxic structural constraints.

Learning Dynamics and Extreme Label Scarcity. Standard evaluations at fixed, abundant label budgets mask the true data-efficiency of message passing. To capture these shifting dynamics, we employ learning curves, which plot a model's generalization performance as a continuous function of the training dataset size [14]. Tracing these entire curves across both homophilic and heterophilic graphs allows us to observe how the relative utility of feature propagation versus topological smoothing shifts dynamically as the node label budget n_l varies.

3 Methodology

This section formalises the node classification problem and details the five-model framework used to isolate the unique contributions of forward-pass topology and explicit backward-pass priors.

3.1 Preliminaries and Problem Formulation

We consider the standard semi-supervised node classification task. Let $G = (V, E)$ be an undirected graph where V is the set of N nodes and E is the set of edges. Each node $v_i \in V$ is associated with a feature vector $\mathbf{x}_i \in R^F$, forming a feature matrix $\mathbf{X} \in R^{N \times F}$. During training, we are provided with ground truth labels for a small subset of nodes $L \subset V$, determined by the label budget n_l per class. The goal is to predict labels for the unlabelled nodes $U = V \setminus L$.

To perfectly align our mathematical framework with our numerical implementation, the graph structure is represented by an augmented adjacency matrix $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I} \in R^{N \times N}$, which incorporates explicit self-loops. The corresponding degree matrix is defined as $\tilde{\mathbf{D}} = \mathrm{diag}(\tilde{\mathbf{A}}\mathbf{1})$. The unnormalised graph Laplacian is constructed as $\mathbf{L} = \tilde{\mathbf{D}} - \tilde{\mathbf{A}}$.

To evaluate how these models scale under data constraints, we track their performance using a label-budget learning curve [14], which evaluates classification accuracy as a function of n_l while keeping the underlying network architectures fixed.

3.2 Five-Model Evaluation Framework

To systematically determine when graph structure is beneficial, redundant, or detrimental, we compare five distinct configurations:

1. Feature-Only Baseline (MLP). The Multilayer Perceptron relies entirely on the node feature matrix $\mathbf{X}$, ignoring the edge structure E, establishing the empirical feature-floor.

2. Prior-Based Baseline (Graph Tikhonov). Graph Tikhonov regularisation ignores node features entirely, relying only on the graph Laplacian $\mathbf{L}$ and the sparse known labels $\mathbf{Y}_0$ (represented as one-hot vectors for labelled nodes, and zero vectors otherwise). The predictions $\hat{\mathbf{Y}} \in R^{N \times C}$ are obtained by minimising the objective:

$$\min_{\hat{\mathbf{Y}}} \|\hat{\mathbf{Y}} - \mathbf{Y}_0\|_F^2 + \lambda \mathrm{Tr}(\hat{\mathbf{Y}}^\top \mathbf{L} \hat{\mathbf{Y}}) \tag{1}$$

where λ controls the strength of the smoothness penalty (fixed at $\lambda = 1.0$). Setting the gradient to zero yields the closed-form solution $\hat{\mathbf{Y}} = (\mathbf{I} + \lambda \mathbf{L})^{-1} \mathbf{Y}_0$.

3. Implicit Topology + Features (GNN). We use ChebNet [4] as our representative spectral GNN, which implicitly integrates structural awareness during the forward pass via Chebyshev polynomial filters.

4 & 5. Explicitly Regularised Models (MLP + Laplacian & GNN + Laplacian). To test the effects of explicit structural constraints, a Laplacian penalty is appended to the standard cross-entropy loss:

$$\mathcal{L}_{reg} = \lambda \frac{1}{|E|} \sum_{(u,v) \in E} \|\mathrm{softmax}(\hat{\mathbf{y}}_u) - \mathrm{softmax}(\hat{\mathbf{y}}_v)\|_2^2 \tag{2}$$

Applying this to the MLP (Model 4) forces a feature-only architecture to optimize for topological smoothness. Applying it to the GNN (Model 5) evaluates the optimization interaction between forward-pass feature mixing and backward-pass penalties.

3.3 Analytical Tracking Metrics

Label Smoothness (Graph Dirichlet Energy). While standard edge homophily calculates the discrete probability of neighbouring nodes sharing a class, we quantify dataset

smoothness using the normalized Dirichlet Energy of the labels. As established by Kalofolias [8], the quadratic form $\mathrm{Tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X})$ defines global signal smoothness on a graph. By expanding this framework to categorical distributions, we convert the labels $\mathbf{Y}$ into one-hot encoded vectors and evaluate their structural friction:

$$\mathcal{E}_{Dirichlet} = \frac{1}{|E|} \sum_{(u,v)\in E} \|\mathbf{Y}_u - \mathbf{Y}_v\|_2^2 \tag{3}$$

This formulation evaluates the exact Laplacian cost function ($\mathrm{Tr}(\mathbf{Y}^\top \mathbf{L}\mathbf{Y})$) scaled by the edge volume. A low Dirichlet Energy indicates a smooth topology where the Laplacian prior serves as an accurate structural anchor, whereas a high value indicates chaotic heterophily where the smoothness prior is fundamentally violated.

Auxiliary Optimization Diagnostic (Gradient L_2 Norm). To inspect how explicit optimization constraints shift the underlying training dynamics of Model 5, we track the L_2 norm of the parameter gradients across training epochs t:

$$\|\nabla \mathcal{L}^{(t)}\|_2 = \sqrt{\sum_{\theta\in\Theta} \left(\frac{\partial \mathcal{L}^{(t)}}{\partial \theta}\right)^2} \tag{4}$$

where Θ represents the complete set of learnable network weights. We employ this diagnostic strictly as an exploratory metric to evaluate relative optimization stability under competing objectives.

4 Experimental Setup

4.1 Datasets and Structural Properties

We evaluate our five-model framework across four established node classification benchmarks spanning distinct structural regimes. To theoretically ground our empirical results, we first quantify the inherent signal properties of each dataset using Label Smoothness (Dirichlet Energy) and Average Node Degree.

As shown in Table 1, Cora and PubMed [13] serve as our sparse, homophilic baselines, exhibiting low Dirichlet Energy (≈ 0.39), mathematically satisfying the classical smoothness prior [8]. Conversely, the Chameleon and Squirrel Wikipedia networks [11] serve as our heterophilic baselines. For these web graphs, we strictly employ the filtered sub-graphs introduced by Platonov et al. [12] to prevent the severe structural data leakage caused by overlapping duplicate nodes in the original releases. Both filtered datasets exhibit severe signal friction (Dirichlet Energy > 1.5), violating the fundamental assumption of Laplacian-based solvers. Crucially, while both are heterophilic, Squirrel is massively denser, possessing an average degree more than twice that of Chameleon.

Table 1: Empirical structural properties of the evaluation datasets.

Dataset	Label Smoothness	Avg. Degree	Topology Type
Cora	0.3801	~ 3.9	Homophilic / Sparse
PubMed	0.3952	~ 4.5	Homophilic / Sparse
Chameleon	1.5279	~ 19.9	Heterophilic / Medium
Squirrel	1.5856	~ 42.3	Heterophilic / Dense

4.2 Architecture and Hyperparameter Setup

To ensure a fair comparison, the neural network models share equivalent capacity constraints. Both the feature-only MLP and the topology-aware GNN (ChebNet) are configured with exactly 2 layers and a hidden dimension width of 64 channels. For ChebNet, we utilise a Chebyshev polynomial filter of order $K = 3$ and apply symmetric normalisation to the graph Laplacian. For explicitly regularised models (MLP+Laplacian, GNN+Laplacian) and the pure Tikhonov solver, the regularisation strength is fixed at $\lambda = 1.0$. This specific value was selected to serve as a robust, balanced prior that mitigates immediate over-smoothing on heterophilic graphs, with the full sensitivity analysis detailed in Appendix A.

All neural models are trained using the Adam optimizer with a learning rate of 0.01 and a weight decay of 5×10^{-4}. To prevent overfitting, we apply dropout with a probability of 0.5 between layers. These optimization configurations were derived from a grid search evaluated on the baseline MLP (see Appendix A) and subsequently standardized across all parameterised architectures to strictly isolate the effects of topological propagation. Training proceeds for a maximum of 200 epochs (100 for the MLP baselines), with an early stopping patience of 50 epochs evaluated on the validation set.

4.3 Experimental Protocol and Reproducibility

The core experimental variable is the label budget n_l, representing the number of training examples provided per class. We dynamically sweep n_l from extreme scarcity ($n_l = 5$) up to data-rich regimes (e.g., $n_l = 1280$ for PubMed).

For rigorous confidence intervals and to account for the high variance of random sampling at low budgets, the homophilic experiments are averaged over 20 random initialisation and sampling seeds. For the heterophilic datasets, we use 10 predetermined structural splits, each evaluated with 2 random initialisation seeds, yielding 20 total runs per budget. For complete reproducibility of our findings, the exact evaluation harness, split logic, and results are saved in the Git repository.

5 Results and Discussion

To evaluate model performance across our defined informational regimes, we report the macro-F1 score (where higher is better). The macro-F1 metric calculates the unweighted mean of the F1 scores across all classes, ensuring that minority classes are not overwhelmed by majority classes in imbalanced test sets.

By analysing learning curves across diverse label budgets, we isolate distinct operational regimes dictated by dataset smoothness and feature richness. In this section, we present the empirical results alongside a quantitative analysis of model variance, and concurrently discuss the theoretical mechanisms driving model behaviour.

Data Scarcity and Evaluation Variance. A prominent feature of our learning curve visualizations (indicated by the shaded standard deviation bands) is the visible variance at low label budgets. This is a well-documented phenomenon in

few-shot graph learning; when the training pool is extremely small (e.g., $n_l = 5$), the specific random sampling of nodes heavily dictates generalization performance. As quantified in Table 2, as the label budget increases, the standard deviation generally contracts (e.g., the Cora GNN standard deviation halves from ± 0.061 to ± 0.029), allowing for a clean visual separation of the models.

However, on heterophilic datasets like Chameleon, the variance bands appear heavily overlapped across the entire budget range. It is crucial to note that this visual overlap is not just a symptom of evaluation instability, but a direct reflection of model convergence: the mean performance floors of the architectures are so identical (all clustering near 0.28) that their standard deviations inevitably intersect.

5.1 Homophilic Regimes and the Feature-Dominance Effect

Figure 1 displays the macro-F1 learning curves for our two homophilic networks. On the left panel (Cora), we observe a clear structural benefit gap: the standard GNN (green) and explicitly regularised models maintain a substantial lead over the feature-only MLP (blue) at moderate and high label budgets. Furthermore, at extreme label scarcity ($n_l = 5$), the topology-only Tikhonov baseline (red) matches the regularised neural models, scoring significantly higher than the standard MLP.

To assess the mathematical significance of these observations, we performed Welch's Two-Sample T-Test [15] at the $n_l = 160$ budget. Welch's test is specifically utilized here because it reliably evaluates differences in means between populations with unequal variances. On Cora, the GNN's lead over the feature-only MLP is highly significant ($t = 34.57$, $p < 0.0001$), as is its lead over the topology-only Tikhonov regularisation ($t = 15.80$, $p < 0.0001$). This empirical evidence provides statistical confidence that the graph structure yields a substantial, non-trivial performance edge under homophily. These observations are consistent with the classical homophily assumption, that connected nodes tend to share labels, suggesting that label propagation acts as a beneficial structural anchor when feature data is scarce.

Conversely, the right panel of Figure 1 (PubMed) presents a radically different dynamic. Despite PubMed being a smooth, homophilic graph, the standard MLP performs nearly identically to the GNN at higher budgets (both converging near ≈ 0.84 macro-F1). The convergence of these models suggests that graph structure may become computationally redundant if the intrinsic node features are highly informative. PubMed utilizes sophisticated TF-IDF vectors of medical abstracts [13], whereas Cora relies on sparse bag-of-words. When feature vectors inherently contain enough discriminative signal to separate classes, the marginal utility of topological message passing, either implicit or explicit, diminishes drastically.

5.2 The Heterophilic Trap and Catastrophic Mixing

Figure 2 illustrates learning curves on the heterophilic web graphs. On Chameleon (left panel), all five models cluster tightly together, flatlining at a mathematical floor of approximately $0.28 - 0.30$ macro-F1. No model, topology-aware or feature-only, escapes this boundary, and explicit Laplacian regularisation fails to degrade the neural architectures below the MLP baseline.

Welch's Two-Sample T-Test at the $n_l = 160$ budget confirms that the apparent performance differences on Chameleon are mathematically negligible. The performance delta between the GNN and the MLP is not statistically significant ($t = 1.32$, $p = 0.196$), nor is the difference between the GNN and the pure Tikhonov solver ($t = 1.22$, $p = 0.230$). Taken together with our homophilic tests, these metrics suggest that graph structure provides a tangible performance edge primarily when the network configuration aligns with semantic class designations.

One plausible explanation for this joint convergence is that the structural web features are fundamentally uninformative for the target task. Furthermore, the fact that explicitly regularised models ('mlp_reg', 'gnn_reg') maintain performance equivalent to the MLP baseline rather than degrading demonstrates optimization resilience. This behavior is consistent with the Adam optimizer balancing competing loss terms, adaptively down-weighting the highly restrictive structural constraint to prioritize cross-entropy minimization.

However, on the Squirrel dataset (Figure 2, right panel), we observe a strong divergence. The Tikhonov baseline experiences a severe performance degradation, dropping to 0.18 macro-F1, significantly worse than the feature-only MLP (0.24). Welch's Two-Sample T-Test at the $n_l = 160$ budget confirms this collapse is highly statistically significant ($t = -7.77$, $p < 0.001$). Furthermore, comparing the neural architectures reveals that the GNN performs significantly worse than the MLP ($t = -3.57$, $p = 0.001$), indicating that implicit forward-pass mixing is actively detrimental here.

While both Chameleon and Squirrel possess identically high Dirichlet Energy (violating the smoothness prior), Squirrel has an average degree more than two times higher (~ 42.3, as detailed in Table 1). When the non-parametric Tikhonov solver attempts to analytically diffuse labels across 42 contradicting, heterophilic neighbours simultaneously, it is highly susceptible to over-smoothing. The solver is forced into uniform probability predictions, resulting in a dramatic loss of accuracy. This suggests that heterophily alone is not the sole factor limiting classical solvers; rather, it is the combination of heterophily and dense connectivity.

5.3 Training Dynamics and Objective Tension

To investigate how explicitly regularised models balance cross-entropy against topological penalties, we tracked the total gradient L_2 norm as an exploratory diagnostic. As observed in Figure 3, the explicitly regularised model on Cora establishes a higher, yet relatively smoother, gradient floor compared to the late-stage jitter of the unconstrained model.

However, because the regularised model's total gradient comprises both cross-entropy and Laplacian contributions, a larger total gradient norm does not inherently prove optimisation stabilisation. Instead, it may simply indicate that the regularisation term remains actively in tension with the primary classification loss. The total norm alone does not reveal

Table 2: Quantitative comparison of mean macro-F1 scores ($\pm$ one standard deviation) at extreme scarcity ($n_l = 5$) and moderate availability ($n_l = 160$). As expected, the standard deviation heavily contracts as the label budget scales, resolving early-stage variance.

Dataset	Model	$\mathbf{n_l = 5}$	$\mathbf{n_l = 160}$
Cora	Feature-Only (MLP)	0.121 ± 0.044	0.547 ± 0.025
	Topology-Only (Tikhonov)	0.219 ± 0.077	0.706 ± 0.018
	GNN (ChebNet)	0.118 ± 0.061	0.756 ± 0.029
Chameleon	Feature-Only (MLP)	0.175 ± 0.032	0.282 ± 0.036
	Topology-Only (Tikhonov)	0.191 ± 0.045	0.284 ± 0.032
	GNN (ChebNet)	0.166 ± 0.054	0.295 ± 0.027

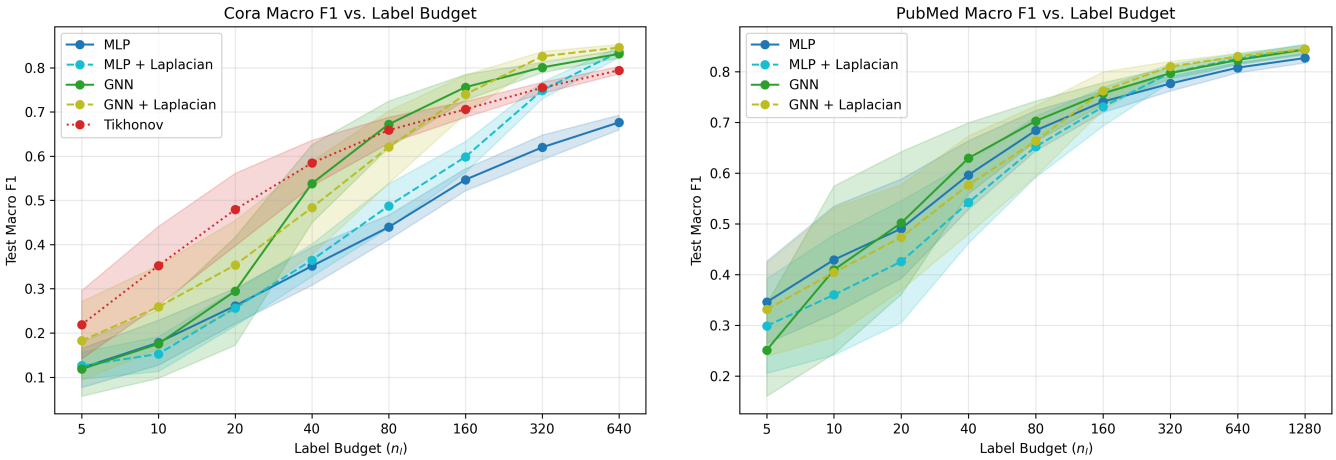


Figure 1: Test macro-F1 learning curves (higher is better) for homophilic datasets Cora (left) and PubMed (right). Curves show the mean over $K = 20$ repetitions per budget, with shaded bands indicating one standard deviation. A clear structural benefit gap is visible on Cora, whereas PubMed exhibits feature-dominance at higher budgets.

whether these two objectives are aligned or mathematically conflicting.

To explicitly disentangle this tension, we performed a secondary diagnostic—tracking the cosine similarity between the isolated cross-entropy and Laplacian gradient vectors across all datasets. We provide these extended diagnostics and remarks in Appendix B for interested readers. Ultimately, these gradient behaviours indicate that explicit Laplacian regularisation fundamentally alters the optimization trajectory, though its utility remains strictly dependent on the underlying structural homophily.

5.4 Threats to Validity

While our controlled experiments isolate the effects of label budgets and topology across four distinct datasets, we acknowledge specific threats to validity. Regarding *external validity*, while Cora, PubMed, Chameleon, and Squirrel effectively contrast sparse homophily with dense heterophily, they do not capture the full diversity of real-world topologies, such as massive scale-free networks, temporal graphs, or multi-relational datasets. Regarding *internal validity*, we restricted the neural models to a fixed architectural capacity to ensure an exact, isolated comparison of topological mechanisms. Dynamically scaling model capacity or employing architecture search specifically tuned for each individual label budget could potentially shift the exact crossover points where feature-dominance or structural degradation occurs.

5.5 Practical Recommendations

Based on our multi-regime empirical findings, we offer actionable recommendations for practitioners deploying models on graph-structured data:

1. **Extreme Scarcity in Smooth Graphs:** When labels are severely limited on homophilic graphs (e.g., Cora), practitioners should deploy simple Laplacian-regularised MLPs or classical label propagation (Tikhonov). These approaches act as structural anchors, avoiding the overfitting risks of deep GNNs.
2. **Feature-Dominant Environments:** If node attributes are highly descriptive (e.g., PubMed text embeddings), standard MLPs are highly efficient. The computational overhead of graph message passing yields diminishing returns and can be safely bypassed.
3. **Dense Heterophilic Networks:** In chaotic structural environments (e.g., Squirrel), pure analytical solvers should be avoided to prevent over-smoothing. A well-tuned, feature-only MLP remains the safest and fastest baseline, as implicit or explicit topological smoothing provides no discriminative advantage.

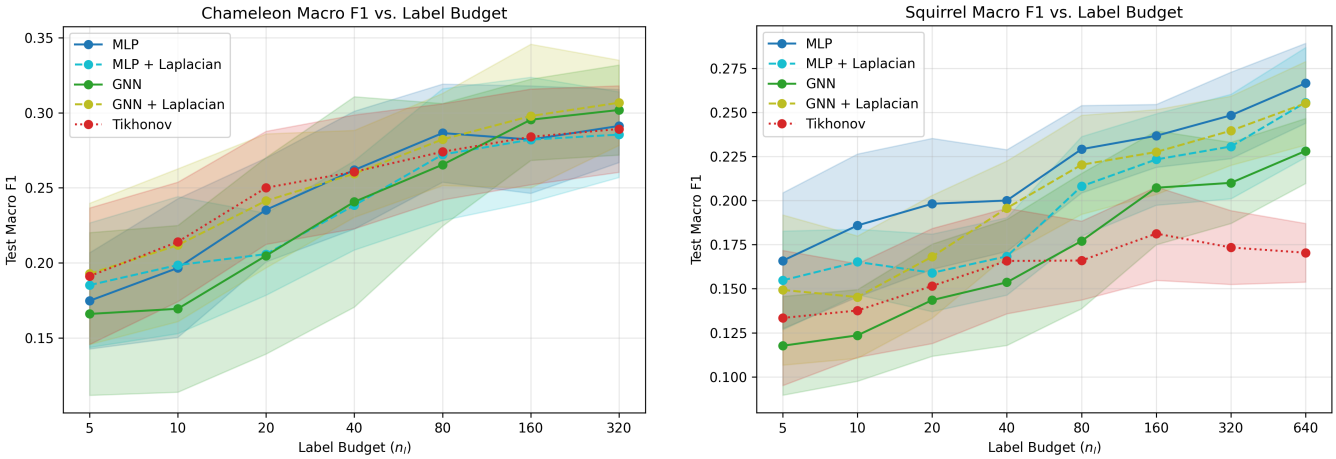


Figure 2: Test macro-F1 learning curves (higher is better) for heterophilic web graphs Chameleon (left) and Squirrel (right). The plots reveal how heavily interconnected heterophilic networks suppress the utility of spatial message passing, forcing architecture variants to converge to or below the feature-only baseline.

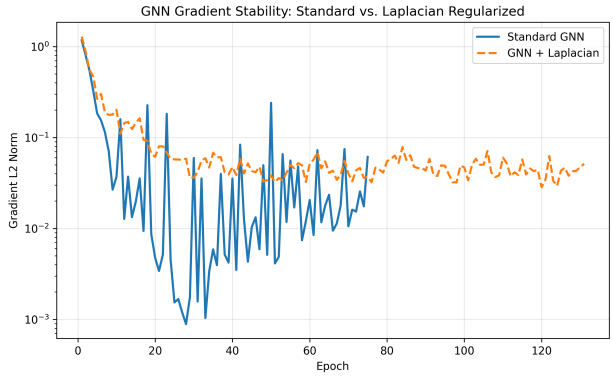


Figure 3: Gradient L_2 norm tracking up to 200 epochs (early-stopped) on Cora ($n_l = 20$), contrasting the optimization stability of the standard GNN against the Laplacian-regularised GNN.

6 Responsible Research

Use of Artificial Intelligence. In accordance with the TU Delft Guidelines on the use of Generative AI, the authors disclose that Google's Gemini large language model was utilized as an assistive technology during this project. Specifically, the model was employed to generate initial code boilerplate and repository infrastructure components, as well as to assist with manuscript proofreading, LaTeX structure formatting, and refining grammatical flow. All core experimental designs, empirical evaluation logic, mathematical derivations, and theoretical interpretations were independently developed and executed by the human authors. The authors have reviewed, verified, and edited all outputs and take full responsibility for the technical accuracy, scientific conclusions, and academic integrity of the final work.

Reproducibility Statement. Ensuring that our experimental results can be reliably reproduced by other researchers is a cornerstone of this work. To mitigate the variance often associated with deep learning on graphs, we implemented a rigorous seed management strategy. All random number generators across Python's `random` module, NumPy, and PyTorch [10] (both CPU and GPU) were strictly initialised using a master seed of 42. Furthermore, PyTorch was configured to enforce deterministic algorithms and disable CuDNN benchmarking (`torch.backends.cudnn.deterministic = True`, `torch.backends.cudnn.benchmark = False`). Our entire evaluation harness, including data loading, model initialisation, and the training loop, is completely deterministic given the seed. We provide the full source code in a public repository, with the exact environment dependencies, including pinned versions of PyTorch and PyTorch Geometric [5], documented in a `requirements.txt` file to ensure identical execution environments.

FAIR Data Statement. In alignment with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, this research relies exclusively on public, well-established benchmark datasets. The Cora, PubMed, Chameleon, and Squirrel datasets are openly accessible and integrated directly via the PyTorch Geometric data loaders, ensuring they are easily findable and accessible to the wider community. We provide our custom data split logic and preprocessing transformations as open-source scripts, ensuring interoperability. The transparency of our data pipeline guarantees that the raw data and our derived experimental splits remain highly reusable for future benchmarking efforts in the graph machine learning community.

7 Conclusions and Future Work

This study evaluated learning curves across five distinct architectural modes for a semi-supervised node classification task. By contrasting implicit forward-pass architectures (ChebNet, MLP) against explicitly regularised backward-pass models and a pure topological baseline (Tikhonov), we aimed to determine when graph structure acts synergistically, redundantly, or detrimentally. Our headline finding is that the utility of graph topology is not universal; rather, it is strictly

bounded by dataset signal smoothness, graph density, and intrinsic feature richness.

In sparse, homophilic environments (Cora), graph structure and node features synergise effectively, and classical label propagation serves as a critical structural anchor at extreme label scarcity. However, if node attributes are inherently highly descriptive (the Feature-Dominance Effect observed on PubMed), complex topological message passing yields diminishing returns and becomes largely redundant.

Furthermore, by quantifying dataset homophily via Dirichlet Energy and tracking gradient stability, we isolated the limitations of explicit topological priors. On heterophilic web graphs, uninformative structural node features trap neural networks at a baseline performance floor. In dense heterophilic networks (Squirrel), enforcing classical smoothness priors induces severe over-smoothing, causing analytical solvers to experience significant performance degradation below this feature-floor. Neural architectures survive these highly restrictive structural constraints primarily through optimization resilience: adaptively down-weighting the explicitly regularised prior. Finally, our gradient tracking suggested that applying explicit Laplacian regularisation to a spectral GNN introduces algorithmic redundancy, damping optimization jitter but yielding no additional discriminative power over the implicit forward pass.

To overcome these limitations in heterophilic and dense environments, future work should move beyond strictly low-pass spectral filters and static homophilic priors. One concrete direction is the deployment of Adaptive Spectral Filters (such as GPR-GNN) [3, 7], which can learn to extract high-frequency signals when local neighbourhoods hold disparate labels. Additionally, exploring Signed Graph Regularisers, where negative edge weights explicitly repel the representations of dissimilar neighbours, could provide a more appropriate inductive bias for heterophily than the standard Laplacian assumption. Finally, developing dynamic architectures that automatically estimate local Dirichlet Energy to adaptively toggle topological constraints could mitigate over-smoothing while maximising data-efficiency across unknown graph topologies.

References

[1] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR)*, 2014.

[3] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network, 2021.

[4] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering, 2017.

[5] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric, 2019.

[6] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 40(3):52–74, 2017.

[7] Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation, 2022.

[8] Vassilis Kalofolias. How to learn a graph from smooth signals, 2016.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[11] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks, 2020.

[12] 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?, 2024.

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

[14] Tom Viering and Marco Loog. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2023.

[15] B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[16] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

[17] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Generalizing graph neural networks beyond homophily. *CoRR*, abs/2006.11468, 2020.

A Hyperparameter Tuning and Sensitivity Analysis

To ensure the integrity of our comparative learning curves, we conducted rigorous hyperparameter tuning on both our

feature-only baseline and our topology-only baseline. This guarantees that any observed performance gaps are due to the structural regimes (implicit vs. explicit topology) rather than under-optimized baselines.

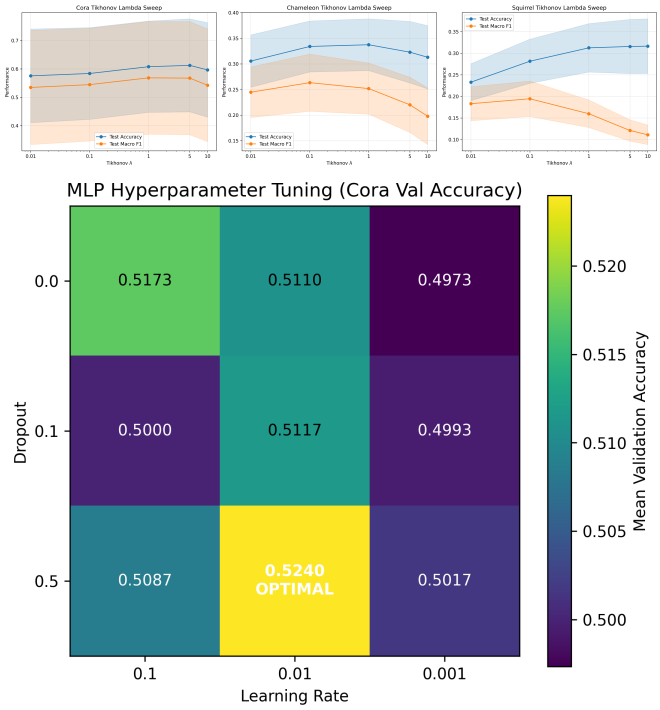


Figure 4: **Left:** Sensitivity analysis of the Tikhonov regularisation strength (λ) on Cora and Chameleon. **Right:** MLP validation accuracy grid search over dropout and learning rate configurations.

A.1 Tikhonov Regularisation Sweep (λ)

The regularisation parameter λ controls the strictness of the explicit topological smoothness prior. To determine a fair, universal value for all regularised models (pure Tikhonov, MLP_REG, and GNN_REG), we swept λ across a logarithmic scale.

As shown in the left panel of Figure 4, the datasets react to the regularisation strength in diametrically opposed ways due to their structural homophily:

- **Cora (Homophilic):** Performance continues to climb as λ increases, peaking around $\lambda = 5.0$. The smooth nature of the citation network heavily rewards strict topological constraints.
- **Chameleon (Heterophilic):** Performance degrades rapidly for any $\lambda > 1.0$. Because the graph is heterophilic, enforcing a strict smoothness prior forces the model to heavily blend conflicting classes, resulting in over-smoothing and a rapid drop in accuracy.

Conclusion: Rather than hyper-optimizing λ per dataset (which would obscure the fundamental structural dynamics we aimed to measure), we fixed $\lambda = 1.0$ for all primary experiments. This value serves as a robust, balanced prior: strong enough to provide a clear structural anchor on sparse homophilic graphs, but not so overwhelming that it instantly triggers catastrophic mixing on heterophilic graphs before the neural optimizer can adapt.

A.2 Feature-Only Baseline (MLP) Grid Search

A common pitfall in Graph Neural Network literature is the under-tuning of the feature-only MLP baseline, which artificially inflates the perceived value of topological message passing. To prevent this, we performed a comprehensive grid search over the MLP's learning rate and dropout probability, evaluating the combinations based on validation macro-F1 scores.

As illustrated in the right panel of Figure 4, the optimization landscape reveals a clear preference for a moderate learning rate (0.01) paired with a relatively high dropout rate (0.5).

- **Learning Rate:** Slower learning rates (0.001) failed to converge within our early-stopping patience window, while aggressive rates (0.05) resulted in severe gradient instability and sub-optimal local minima.
- **Dropout:** Because the benchmark datasets (especially at low label budgets like $n_l = 20$) are highly susceptible to overfitting, standardizing a dropout probability of 0.5 proved essential for forcing the neural networks to learn generalized feature representations rather than memorizing the sparse training split.

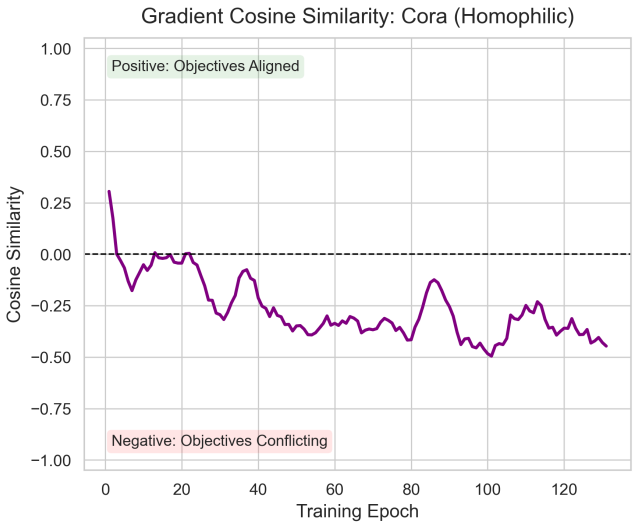
Conclusion: The optimal configuration (Learning Rate $= 0.01$, Dropout $= 0.5$) was subsequently locked and applied universally across all parameterised neural models (MLP, ChebNet, MLP_REG, GNN_REG) to ensure a strictly controlled ablation of architectural capabilities.

B Extended Diagnostic: Gradient Cosine Similarity

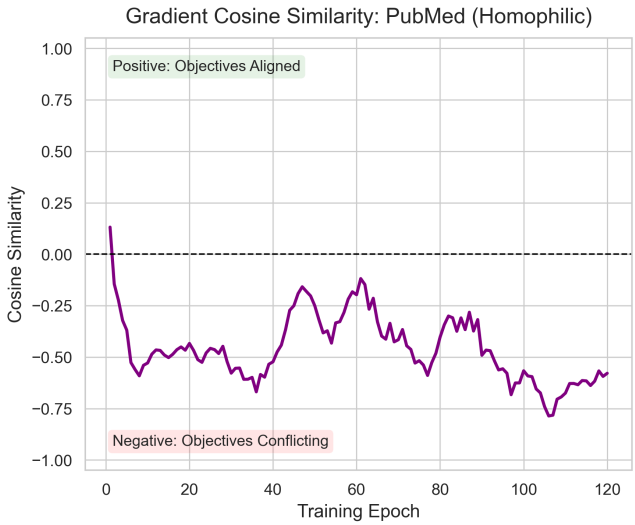
Building upon the gradient L_2 norm observations in Section 5.3, we provide a more in-depth diagnostic to disentangle the total gradient into its constituent parts. Because a high total gradient norm can mask competing objectives, we tracked the cosine similarity between the isolated cross-entropy gradient vector and the Laplacian penalty gradient vector at each epoch. A negative cosine similarity indicates that the explicitly enforced topological prior is mathematically opposing the parameter updates required by the classification labels.

As illustrated in Figure 5, the cosine similarity is universally negative across all datasets, reflecting the expected baseline behaviour of a regulariser fighting the primary loss to prevent overfitting. However, the trajectory and magnitude of this tension vary distinctly based on structural homophily:

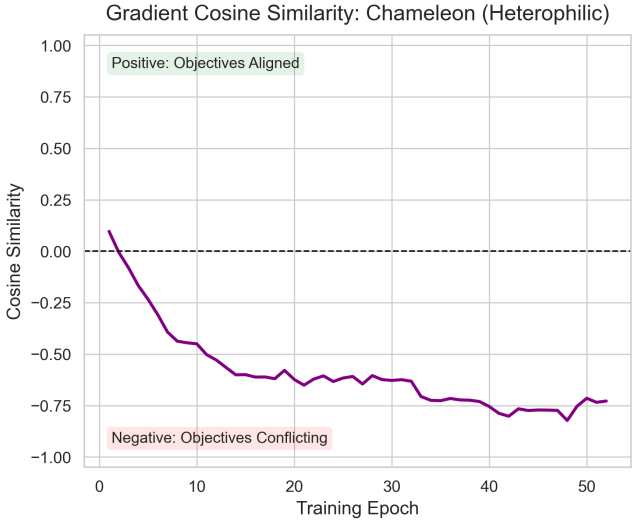
- **Homophilic Regimes (Cora, PubMed):** The cosine similarity exhibits high variance (a highly "jittery" trajectory) and fluctuates in a moderately negative range. This suggests a dynamic negotiation where the regulariser restrains the optimizer but does not persistently force it in the exact opposite direction of the semantic signal.
- **Heterophilic Regimes (Chameleon, Squirrel):** The cosine similarity drops rapidly and remains heavily suppressed (closer to -1.0), exhibiting a much smoother,



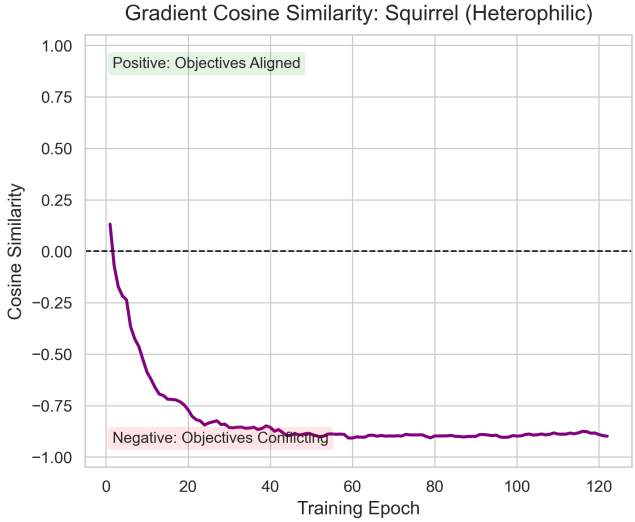
(a) Cora (Homophilic)



(b) PubMed (Homophilic)



(c) Chameleon (Heterophilic)



(d) Squirrel (Heterophilic)

Figure 5: Cosine similarity between the cross-entropy and Laplacian gradients during the training of the regularised GNN. These plots act as an extended diagnostic to observe objective alignment across varying structural regimes.

persistent trajectory of opposition. This indicates that on networks with high signal friction, the explicitly enforced smoothness prior remains in constant, direct conflict with the cross-entropy objective.

These visualisations are provided as an exploratory look into objective tension under varying structural regimes and are intended to supplement the primary learning curve analysis.