

Exploratory Paper Analysis: Why Does Your Graph Neural Network Fail on Some Graphs?

Abstract

Graph Neural Networks (GNNs) often succeed on benchmark graphs yet fail on others in ways that are not fully explained by architectural narratives such as over-smoothing or over-squashing. The target paper reframes this problem through exact generalisation error, showing that alignment between graph structure, node features, and labels is the key driver of when a model generalises. This write-up first traces the backward-thinking context that motivated this reframing and then proposes a forward-thinking direction: predictive model auditing as a pre-training diagnostic. We complement the proposal with minimal empirical probes (synthetic sweeps and two real datasets) that illustrate how homophily and feature-signal strength align with performance, supporting the audit idea.

Backward-thinking: Why this paper was needed

Problem setting. GNNs are a standard tool for graph-structured learning, yet performance varies drastically across graph families. This variability is surprising because many GNNs share similar expressivity bounds and are trained in comparable ways, but they still behave inconsistently across datasets. The field has historically explained failures with architectural narratives: over-smoothing in deep message passing, over-squashing under long-range dependencies, or limits of the 1-WL expressivity class. These explanations are useful, but they do not predict why two closely related GNNs can diverge sharply in practice.

Why this matters. When a model fails without a clear diagnostic, practitioners either over-engineer architectures or chase hyperparameters without understanding why performance is fragile. For research, the consequence is worse: benchmark-driven narratives can mislead the community into believing a new architecture is universally superior when its success is tied to specific data regimes. The paper positions itself against this pattern, asking for a principled quantity that can explain failure across multiple model families, not just one architecture.

Limits of prior theory. The literature contains generalisation error bounds for GNNs, but they are often loose, tied to a single architecture, or too abstract to diagnose real failures. As a result, researchers have strong but fragmented stories: architectural limitations explain some trends, while empirical benchmarks suggest other trends. The missing link is a principled, model-agnostic quantity that explains when the graph structure helps and when it hurts.

From spatial vs. spectral to alignment. A useful backdrop is the long-standing divide between spatial GNNs (neighborhood aggregation) and spectral GNNs (filters in the Laplacian eigenbasis). Recent work has shown these views can be unified by thinking in terms of frequency response. The target paper leverages this unification: by treating many GNNs as graph filters, it becomes possible to ask whether the label signal lies in frequencies preserved by the model. This is the conceptual bridge that moves the discussion from architecture taxonomy to data-model alignment.

The paper’s reframing. The paper addresses this gap by deriving an *exact generalisation error* for a broad class of GNNs under a transductive, fixed-design setting. The key move is a signal-processing perspective: graph convolutions are interpreted as graph filters, and the alignment between the spectral structure of the graph and the signal in node features/labels becomes the decisive factor. This shifts the question from *which architecture is best* to *when does graph structure provide aligned signal that a GNN can exploit*?

Signal-processing lens. In this view, GNN layers behave like filters in the graph spectral domain. The data signal is decomposed into graph frequencies, and a model’s frequency response determines which parts of the signal are amplified or suppressed. This makes two concepts central: (i) the spectral distribution of the graph itself, and (ii)

whether the target signal is concentrated in frequencies that a given model can preserve. The paper argues that failures often occur not because a model is weak in an abstract sense, but because the graph signal is misaligned with what the model can represent.

Core failure mechanisms highlighted by the paper.

1) *Misalignment*. Even shallow GNNs can fail if the adjacency structure, features, and targets are misaligned. The paper formalises this with a misalignment measure: when the target signal does not lie in the subspace captured by the GNN representation, generalisation error increases. This explains why a simple concatenation baseline can outperform convolution in some regimes.

2) *Heterophily as high-frequency signal*. The paper connects homophily/heterophily to spectral properties: homophilic labels are low-frequency signals, while heterophilic labels are high-frequency. Convolutional GNNs are biased toward low-pass behavior, which makes them struggle when the label signal is high-frequency. This ties failures to the graph’s spectrum rather than only to the model class.

3) *Attention limitations under repeated eigenvalues*. For attention-based models, the paper shows a spectral limitation: when eigenvalues repeat, certain attention mechanisms cannot fully distinguish graph structure, revealing a structural blind spot even for flexible models. This again suggests that graph spectral properties are as important as model design.

Empirical evidence in the paper. The paper uses real datasets to illustrate these mechanisms. In the misalignment analysis, convolution can lose to simple concatenation when the graph and features are poorly aligned, highlighting that graph structure is not universally beneficial. For heterophily, the paper ties poor performance to high-frequency label signals that are suppressed by common GNN filters. The empirical claims reinforce the message: failures are predictable once alignment and spectral structure are quantified.

Why this reframing changes evaluation. If alignment and spectrum are the governing variables, then benchmark success is partly an artifact of dataset selection. In a benchmark suite dominated by alignment-friendly graphs, architectures tuned for low-pass behavior will appear robust. Conversely, in heterophilic or misaligned regimes, the same architectures will appear brittle. This suggests a need for evaluation protocols that cover a broader range of alignment conditions rather than a single, convenient slice of the graph space.

Why benchmarks can mislead. The paper also argues that standard benchmarks (e.g., Cora, CiteSeer) are biased toward alignment-friendly regimes. This means GNNs can appear strong in evaluation while still being fragile in settings with weaker alignment or heterophilic structure. The broader implication is that failure is not an exception; it is the predictable outcome of misalignment between data and graph structure.

Backward-looking synthesis. The historical focus on architectural fixes explains only part of the failure landscape. The paper’s key contribution is to show that *alignment and spectral structure are the controlling variables*. This reframes GNN failures from *model design flaws* to *data-model mismatch*, setting the stage for a forward-looking diagnostic approach.

Forward-thinking: Predictive Model Auditing

Proposal. A natural research direction is to operationalise the paper’s insights as a *pre-training diagnostic*: Predictive Model Auditing. The idea is to compute a failure risk score from graph-structural signals (e.g., homophily, spectral properties, feature-label alignment proxies) before any training occurs. This moves the field from post-hoc explanations toward pre-flight assessments of when a GNN is expected to generalise.

What the audit predicts. The audit is not a single number that replaces training; rather, it is a set of warnings and expectations. If a graph exhibits low homophily and a feature-label signal dominated by high-frequency components, the audit would predict high risk for standard low-pass GNNs and suggest either alternative filters or preprocessing. Conversely, in strongly aligned regimes, the audit predicts that standard convolutional GNNs should perform reliably, which helps avoid unnecessary architectural complexity.

What signals belong in the audit. A practical audit would combine multiple low-cost signals: (i) homophily as a proxy for label smoothness, (ii) simple spectral summaries such as the distribution of Laplacian eigenvalues, and (iii) feature-label alignment proxies (for example, correlations between features and labels on training nodes). Each signal is imperfect, but the combination provides a robust early warning system for failure-prone regimes.

Audit pipeline.

1. **Audit stage:** Compute structural signals from the graph and label distribution to estimate a failure risk score (e.g., inverse homophily as a proxy for high-frequency label signals).
2. **Decision stage:** If risk is high, select architectures or preprocessing steps better aligned with the data (e.g., alternative filters, rewiring, or feature augmentation).
3. **Validation stage:** Stress-test predictions on controlled graph families to verify that the risk score correlates with actual failure rates.

Empirical probes supporting the audit idea. We ran minimal experiments to test whether structural signals predict performance. These are intentionally small-scale, designed to validate directionality rather than to maximize accuracy.

Synthetic homophily sweep. On synthetic SBM graphs, GNN accuracy increases with homophily, while an MLP baseline remains comparatively flat. A simple risk score (1 - homophily) correlates strongly and negatively with GNN accuracy, consistent with the audit hypothesis. This aligns with the paper’s claim that homophily corresponds to low-frequency signals that standard GNNs can exploit.

In our runs, the negative correlation between the risk score and GNN accuracy is consistently strong across model families, while the MLP baseline remains weakly correlated. This separation is exactly what a useful diagnostic should detect: when the graph signal aligns with the task, GNNs benefit; when it does not, the advantage disappears.

Alignment (feature-signal) sweep. Holding graph structure fixed and varying how much label signal appears in node features, model performance improves as feature signal strength increases. This directly supports the misalignment hypothesis: weak feature-label alignment degrades generalisation even when graph structure is unchanged. Importantly, the effect is stronger for GNNs than for the MLP baseline, suggesting that graph structure does not compensate for poor feature-label alignment.

Real datasets. On Cora and CiteSeer, higher homophily corresponds to higher GNN accuracy. The multi-seed results below show consistent gains for GNNs over MLP, with better performance on the more homophilic dataset (Cora). These results align with the paper’s claim that benchmark success is linked to alignment-friendly regimes.

Interpretation of the real-data table. The gap between GNNs and MLP is larger on Cora than on CiteSeer, consistent with stronger alignment. The variance across seeds is modest, indicating the trend is stable rather than accidental. This provides a concrete anchor for the audit: it is not an abstract claim, but one that predicts observable differences on widely used benchmarks.

Dataset	Model	Homophily	Test Acc (mean)	Std
Cora	GCN	0.810	0.7898	0.0102
Cora	GraphSAGE	0.810	0.7664	0.0069
Cora	GAT	0.810	0.7556	0.0258
Cora	MLP	0.810	0.4930	0.0250
CiteSeer	GCN	0.736	0.6340	0.0174
CiteSeer	GraphSAGE	0.736	0.6334	0.0125
CiteSeer	GAT	0.736	0.6380	0.0111
CiteSeer	MLP	0.736	0.4806	0.0061

Why this is enough for a forward-looking claim. These probes are not exhaustive, but they demonstrate the core mechanism: performance tracks alignment-related signals across both synthetic and real graphs. This supports the feasibility of a pre-training audit that predicts when GNNs are likely to fail.

How the audit would be used in practice. In a research workflow, the audit could be applied before model development to select a family of architectures likely to succeed. In an industrial workflow, it could screen datasets for risky alignment regimes and recommend whether to invest in graph-specific modeling at all. In both settings, the audit provides a principled rationale for why a baseline is chosen and when more complex modeling is justified.

Roadmap and limitations. A next step is to build a standardized stress suite that varies homophily, degree regimes, and spectral properties, then calibrate a multi-signal risk score against observed failures. The audit should be transpar-

ent about what it captures (alignment and spectrum) and what it does not (label noise, task-specific complexity). The goal is not to replace training, but to avoid blind reliance on benchmarks that may hide failure modes.

Coherent closing. The backward-looking story reframes failure as misalignment; the forward-looking proposal makes that reframing actionable. Predictive model auditing is a direct bridge from theory to practice: it uses alignment and spectral signals to predict where GNNs will generalise and where they will fail, before expensive training begins.

References

- Ayday, N., Sabanayagam, M., & Ghoshdastidar, D. (2025). *Why Does Your Graph Neural Network Fail on Some Graphs? Insights from Exact Generalisation Error*.
- Craven, M., et al. (1998). *Cora*.
- Giles, C. L., et al. (1998). *CiteSeer*.