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. I 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.

1 Backward-thinking: Why this paper was needed

This analysis examines the paper "Why Does Your Graph Neural Network Fail on Some Graphs? Insights from Exact Generalisation Error" [citation]. The paper addresses the question why performance varies across graphs and derives an exact generalisation error framework in a transductive fixed design setting, using a signal processing lens that treats GNNs as graph filters. It positions exact generalisation error and alignment as the central quantity for explaining success and failure across architectures. This analysis uses the paper to ground the backward looking discussion, to identify the problem that led to producing this paper, to clarify the limits of prior explanations, and to motivate a forward looking diagnostic proposal. It then proposes a diagnostic tool and evaluates it on experiments, including synthetic sweeps and real dataset checks, in order to test its predictions.

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.

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.

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?* This reframing also connects spatial and spectral views: many architectures can be described by their frequency response, so failures can be traced to misaligned signal frequencies rather than to a specific model family.

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 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.

2 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 summaries, and 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.

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.

Experimental design choices

I include GCN, GraphSAGE, GAT, GPRGNN, and MLP to span the major model families that appear in practice. GCN represents a standard low pass convolutional baseline, GraphSAGE provides a different aggregation rule, GAT represents attention based architectures, GPRGNN adds a learned spectral response that can emphasize a wider range of frequencies, and MLP provides a feature only baseline. This set allows me to test whether the diagnostic separates graph aware models from a graph agnostic model across diverse architectures.

I use a stress suite of synthetic families to test the diagnostic across controlled axes of variation. SBM with balanced and imbalanced communities provides direct control of homophily and class balance, LFR style graphs add more realistic community structure with tunable mixing, and degree skew graphs test robustness under heavy tailed degree distributions. These families cover homophily control, community realism, and degree heterogeneity in a way that a single generator cannot.

I add Squirrel as a real dataset because it is a widely used heterophily benchmark and it is discussed in the target paper. It provides a realistic setting where low pass biases can fail, which makes it a direct test of the diagnostic beyond homophilic benchmarks like Cora and CiteSeer.

I define the audit score from homophily, a feature label alignment proxy, and a spectral summary such as lambda max. Homophily captures label smoothness, alignment measures how much label signal is represented in features, and the spectral term captures how the graph structure can affect filtering behavior. Combining them turns the diagnostic into a multi signal predictor rather than a restatement of homophily alone.

Empirical probes supporting the audit idea

I ran a stress suite of experiments to test whether structural signals predict performance. The suite spans multiple synthetic graph families with 500 nodes and several seeds, and it includes GPRGNN alongside GCN, GraphSAGE, GAT, and an MLP baseline. The goal is not to maximize accuracy, but to test whether the audit signals hold across families rather than only within a single generator.

Synthetic stress suite. Across SBM balanced and imbalanced, LFR, and degree skew graphs, accuracy increases with homophily for graph-aware models while the MLP baseline remains comparatively flat. The multi-signal risk score is strongly negatively correlated with test accuracy across families, and it is comparable to or slightly stronger than homophily alone for some models such as GAT and GCN, though not uniformly for all models. This supports the audit idea while also showing that better weighting of signals could further improve predictive strength.

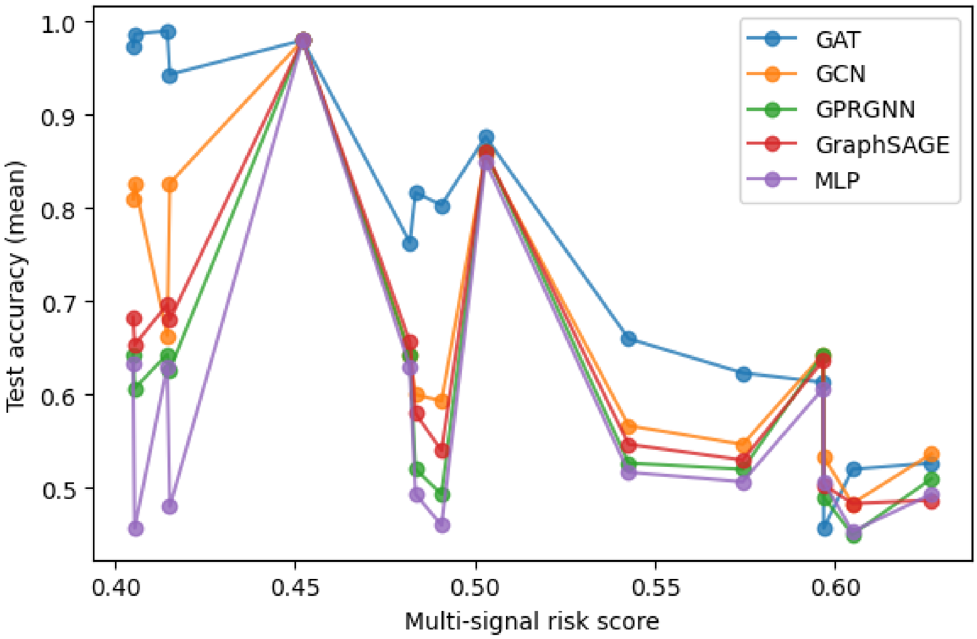

Figure 1: Stress suite results. Test accuracy versus the multi-signal risk score across synthetic families, with lines connecting points within each model. Higher risk is associated with lower accuracy for graph-aware models.

Alignment (feature-signal) sweep. Holding graph structure fixed and varying how much label signal appears in node features, GAT and GCN remain high and stable, GraphSAGE and GPRGNN show mid-range performance, and the MLP baseline stays near chance. The trend is not perfectly monotonic across feature strengths, but the overall pattern supports the misalignment hypothesis: weak feature-label alignment degrades generalisation even when graph structure is unchanged.

Real datasets. On Cora and CiteSeer, which are more homophilic, GNNs consistently outperform MLP and GPRGNN performs best among the tested models. On Squirrel, a heterophily benchmark with much lower homophily, overall accuracy drops and the ranking shifts, with GAT and GraphSAGE outperforming GCN. This aligns with the paper's claim that benchmark success is linked to alignment-friendly regimes and highlights failure modes under heterophily.

Interpretation of the real-data table

The gap between graph-aware models and MLP is largest on Cora, smaller on CiteSeer, and smallest on Squirrel, which is consistent with the homophily ordering. The variance across seeds is modest, indicating the trend is stable rather than accidental. The ordering shift on Squirrel provides a concrete stress test for the audit, showing that model choice matters most when alignment is weak.

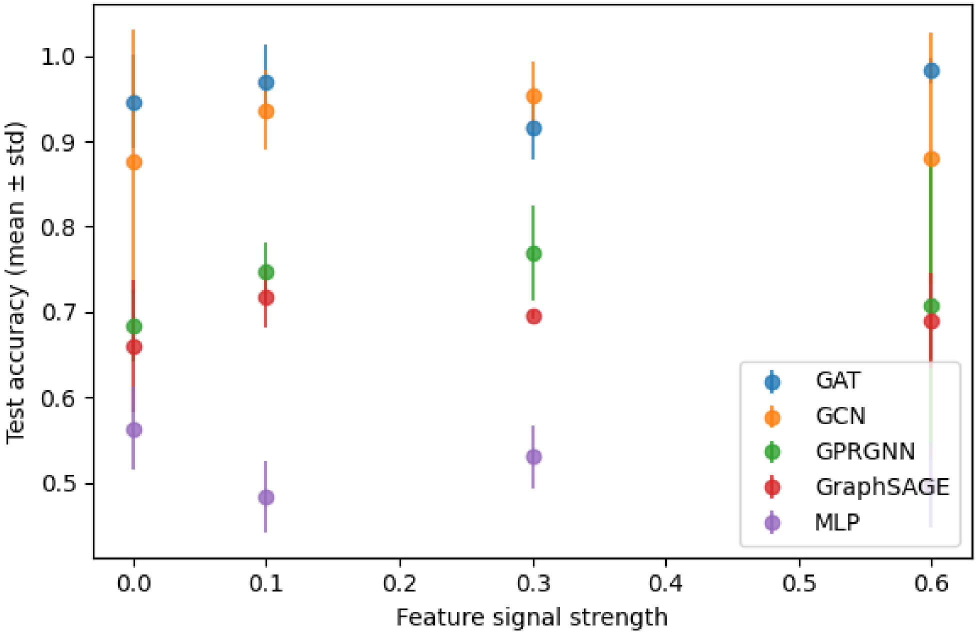

Figure 2: Alignment sweep. Test accuracy versus feature-signal strength with graph structure held fixed.

Dataset	Model	Homophily	Test Acc (mean)	Std
Cora	GPRGNN	0.810	0.8050	0.0041
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	0.4932	0.0247
CiteSeer	GPRGNN	0.736	0.6434	0.0055
CiteSeer	GAT	0.736	0.6380	0.0111
CiteSeer	GCN	0.736	0.6340	0.0174
CiteSeer	GraphSAGE	0.736	0.6334	0.0125
CiteSeer	MLP	0.736	0.4806	0.0061
Squirrel	GAT	0.222	0.4720	0.0293
Squirrel	GraphSAGE	0.222	0.4507	0.0095
Squirrel	GCN	0.222	0.3925	0.0082
Squirrel	GPRGNN	0.222	0.3305	0.0065
Squirrel	MLP	0.222	0.3228	0.0145

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 multiple synthetic families and real datasets, including a heterophily benchmark. The stress suite shows that the audit signal generalises beyond a single graph generator, and the real-data results show that it highlights the regimes where GNNs are most fragile. This supports the feasibility of a pre-training audit that predicts when GNNs are likely to fail.

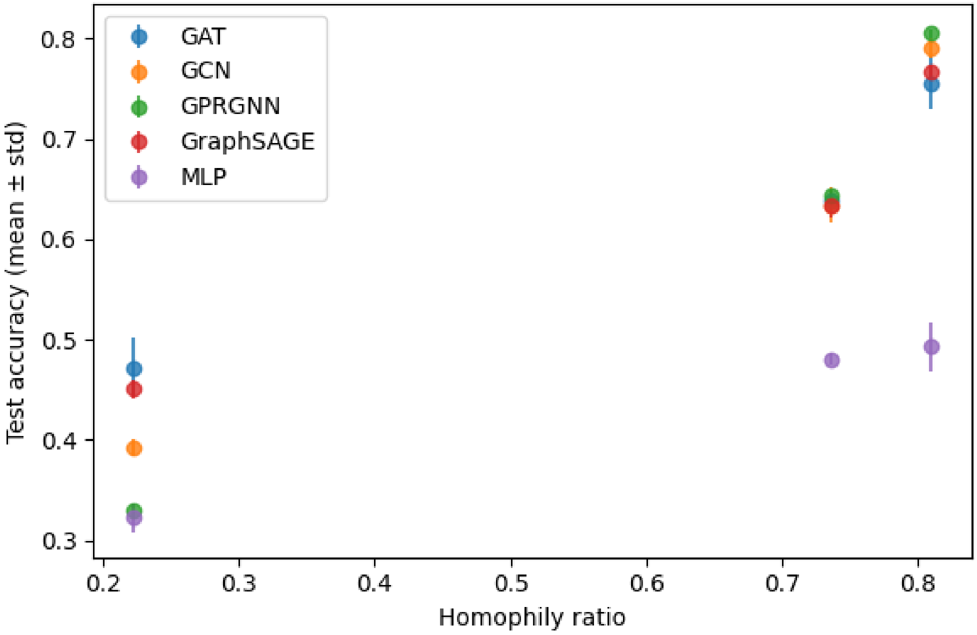

Figure 3: Real datasets. Homophily versus accuracy for Cora, CiteSeer, and Squirrel.

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 transparent 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.

3 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*.