Operational Feature Fingerprints of Graph Datasets via a White-Box Signal-Subspace Probe

Anonymous authors
Paper under double-blind review

Abstract

Graph neural networks achieve strong node-classification performance, but learned message passing entangles ego features, neighborhood smoothing, high-pass graph differences, class geometry, and classifier-boundary effects inside opaque representations. This makes it difficult to determine why nodes are classified as they are, and which graph-learning mechanisms are useful, harmful, or necessary for a given dataset. We propose WG-SRC (*White-box Graph Signal–Subspace Residual Classifier*), a white-box signal-subspace probe for prediction and graph dataset diagnosis. WG-SRC replaces learned message passing with an explicit, named graph-signal dictionary containing raw features, row- and symmetric-normalized low-pass propagation, and high-pass graph differences. It then combines Fisher coordinate selection, class-wise PCA subspaces, closed-form multi-α ridge classification, and validation-based score fusion. Because every signal block and decision module is explicit, the fitted scaffold produces both predictions and an operational fingerprint over raw-feature, low-pass, high-pass, class-geometric, and ridge-boundary mechanisms. Across six node-classification datasets, WG-SRC remains competitive with aligned reproduced baselines and achieves a positive average gain under matched repeated splits. Its fingerprints distinguish low-pass-dominated Amazon graphs, mixed high-pass and class-geometrically complex Chameleon behavior, and raw- or boundary-sensitive WebKB graphs. Aligned interventions further show that these fingerprints are operational: they identify when high-pass blocks behave like removable noise, when graph-derived or raw signals should be preserved, and when ridge-type boundary correction matters. Additional fixed black-box component probes further show that measured dataset fingerprints organize architectural behavior across multiple black-box families: different measured dataset conditions repeatedly favor different inductive biases. Thus, WG-SRC serves both as a functioning white-box classifier and as a dataset-fingerprinting probe, enabling fingerprint-conditioned analysis of how black-box graph-model components behave under different measured dataset conditions.

1 Introduction

Graph neural networks learn by repeatedly aggregating features over edges and applying trained transformations. This recipe is powerful, but it hides several mechanisms behind parameters: whether a prediction is driven by ego attributes, one-hop smoothing, two-hop return structure, high-pass differences between a node and its neighborhood, or a final discriminative boundary is often unclear. This opacity is especially problematic on heterophilic or mixed-homophily graphs, where naive smoothing can hurt because neighbors need not share labels (Pei et al., 2020; Zhu et al., 2020; Lim et al., 2021).

This paper starts from a different premise. Instead of learning a hidden message-passing representation, we explicitly build the graph signals that a shallow graph model might exploit, and classify them using linear-algebraic modules whose behavior is inspectable. The resulting method, WG-SRC (*White-box Graph Signal–Subspace Residual Classifier*), constructs a named graph signal dictionary, selects discriminative coordinates, fits class-wise PCA subspaces, fits a closed-form ridge boundary, and fuses the two scores by validation. In this sense, the method is deliberately built around subspace geometry and controlled low-rank

structure: the same explicit dimension-reduced objects are used both to make decisions and to analyze which graph mechanisms a dataset is using.

The central point is that WG-SRC is a graph-dataset probe implemented through a functioning white-box predictor. After validation selection, the fitted scaffold provides both predictions and a mechanism readout: because every signal block and decision module is named, the same fitted variables define raw-feature reliance, low-pass propagation reliance, high-pass sensitivity, class-subspace complexity, and ridge-boundary dependence. We use *white-box scaffold* for the fitted predictive model and *dataset probe* for the same scaffold when it is used as a measurement instrument.

This view leads to three empirical requirements. First, the scaffold must be predictively functional, so that its readout is produced by a model that has captured useful dataset structure. Second, the fingerprint should be operational: datasets assigned to different signal or decision regimes should respond differently to aligned white-box interventions. Third, the measured dataset condition should organize behavior beyond the WG-SRC scaffold itself, as tested by fixed black-box component probes. The main text follows this chain through predictive validity, mechanistic interventions, dataset fingerprints, and fingerprint-conditioned black-box probing.

Appendix A reports paired random-split stability, Appendix B provides additional atlas views, and Appendix C gives the full black-box component tables.

Contributions. We make three contributions.

1. We introduce WG-SRC, a white-box graph classifier that replaces learned message passing with an explicit multi-hop graph-signal dictionary and replaces hidden representation layers with class-wise PCA subspaces, closed-form multi-alpha ridge regression, and validation-based score fusion.
2. We show that the fitted variables of the same classifier induce an operational graph-dataset fingerprint. The fingerprint aggregates node-level raw-feature, low-pass, high-pass, class-geometric, and boundary-based readouts into dataset-level summaries of signal composition, class-subspace complexity, PCA–Ridge decision structure, and correct-versus-wrong signal shifts.
3. We validate the fingerprint as a diagnostic object. WG-SRC remains competitive under aligned reproduced benchmarks, and its measured fingerprints agree directionally with white-box interventions, signal/error shifts, PCA–Ridge phase structure, and fingerprint-conditioned black-box component probes. These results connect dataset fingerprints to post-evaluation guidance for suppressing noisy high-pass blocks, preserving useful graph-derived or raw signals, strengthening boundary decisions, or improving class-specific subspace modeling.

2 Related Work

Graph neural networks and heterophily. GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), and GAT (Veličković et al., 2018) learn node representations by aggregating local neighborhoods. Heterophilic graphs expose the limitations of pure smoothing. Geom-GCN (Pei et al., 2020), H2GCN-style designs (Zhu et al., 2020), adaptive PageRank filters (Chien et al., 2021), and LINKX (Lim et al., 2021) address non-homophily by changing propagation, decoupling ego and neighborhood features, or using strong simple baselines. WG-SRC follows the decoupling intuition but removes learned hidden layers: it constructs low-pass and high-pass signals explicitly and audits which ones are used.

White-box and subspace learning. PCA and ridge regression are classical tools with transparent objectives (Pearson, 1901; Hoerl & Kennard, 1970). MCR^2 and ReduNet provide a modern white-box perspective on representation learning: classes should occupy structured, discriminative subspaces, and networks can be derived from optimization principles rather than treated as opaque stacks (Yu et al., 2020; Chan et al., 2022; Wang et al., 2024). WG-SRC borrows the subspace viewpoint, but adapts it to graphs by first decomposing node features into named graph signal blocks and then fitting class subspaces in that signal space. Its white-box character therefore comes not only from using explicit graph filters, but also from using subspace geometry, low-rank energy control, and closed-form decision modules as analyzable mathematical objects.

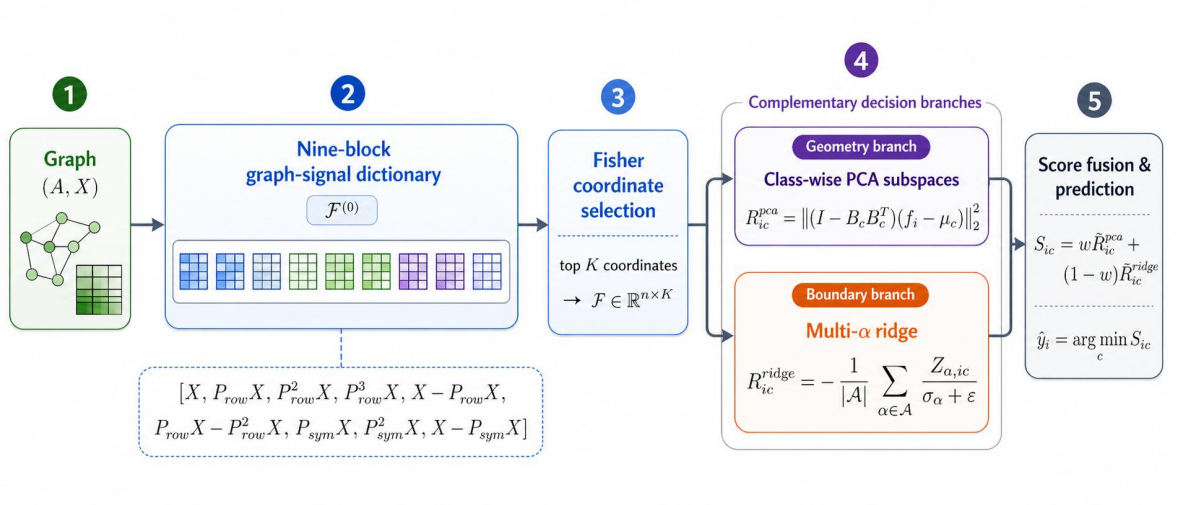


Figure 1: WG-SRC pipeline. The graph is converted into named signal blocks, discriminative coordinates are selected, and prediction is produced by fusing class-subspace residuals with a closed-form ridge boundary.

Transformer-style, latent-graph, and adaptive black-box graph models. Recent black-box graph learners explore several architectural directions beyond classical local aggregation, including simplified global token mixing in SGFormer (Wu et al., 2023), global attention in GOAT (Kong et al., 2023), learned latent graph structure in NodeFormer (Wu et al., 2022), and adaptive expert routing in GNNMoE (Chen et al., 2024). These models can be strong, but their gains are often difficult to attribute to specific graph-learning mechanisms. We therefore use selected black-box families and fixed component ablations as external probes: after WG-SRC measures a dataset fingerprint, these ablations test whether the measured raw-feature, low-pass, high-pass, class-geometric, or boundary-sensitive condition favors corresponding architectural biases.

3 Method

Let $G = (V, E)$ be a graph with feature matrix $X \in \mathbb{R}^{n \times d}$, adjacency matrix A, and labels on a training-node index set $\mathcal{T} \subseteq \{1, \ldots, n\}$. Write $n_{\text{tr}} = |\mathcal{T}|$. WG-SRC has five stages: graph signal construction, Fisher coordinate selection, class subspace fitting, multi-alpha ridge fitting, and score fusion.

3.1 Explicit multi-hop graph signal dictionary

We use the row-normalized transition matrix $P_{\text{row}} = D^{-1}A$ and the symmetric-normalized matrix $P_{\text{sym}} = D^{-1/2}AD^{-1/2}$, following standard normalized graph-operator conventions (von Luxburg, 2007). The block-normalized graph signal dictionary is

$$F^{(0)} = \left[X,\ P_{\text{row}}X,\ P^2_{\text{row}}X,\ P^3_{\text{row}}X,\ X - P_{\text{row}}X,\ P_{\text{row}}X - P^2_{\text{row}}X,\ P_{\text{sym}}X,\ P^2_{\text{sym}}X,\ X - P_{\text{sym}}X\right]. \quad (1)$$

Each block is row-ℓ_2 normalized before concatenation, so $F^{(0)} \in \mathbb{R}^{n \times p}$ with $p = 9d$. After Fisher selection we obtain a coordinate set $\mathcal{S} \subseteq \{1, \ldots, p\}$ with $|\mathcal{S}| = K$, and we write

$$F = F^{(0)}_{[:,\mathcal{S}]} \in \mathbb{R}^{n \times K}, \qquad F_{\text{tr}} = F_{\mathcal{T},:} \in \mathbb{R}^{n_{\text{tr}} \times K}. \quad (2)$$

The downstream PCA, ridge, and atlas computations all use these selected coordinates unless stated otherwise.

3.2 Fisher coordinate selection

For coordinate j of the full dictionary $F^{(0)}$, the Fisher score is

$$q_j = \frac{\sum_{c=1}^{C} n_c(\mu_{c,j} - \mu_j)^2}{\sum_{c=1}^{C} \sum_{i \in \mathcal{T}:\, y_i = c} (F^{(0)}_{ij} - \mu_{c,j})^2 + \epsilon}. \quad (3)$$

The numerator measures between-class separation and the denominator measures within-class scatter. The top K coordinates define the selected set $\mathcal{S}$ in Eq. (2). The value of K is selected by validation.

3.3 Class subspace residuals

For each class c, PCA is fit on the selected training matrix F_{tr} restricted to class c. Let $\mu_c \in \mathbb{R}^K$ be the class center, let $B_c \in \mathbb{R}^{K \times r_c}$ be the orthonormal basis selected by an energy threshold, and let $f_i = F_{i,:}^\top \in \mathbb{R}^K$ denote the selected feature vector of node i. The class-subspace residual score is

$$R_{ic}^{\mathrm{pca}} = \left\| (I - B_c B_c^\top)(f_i - \mu_c) \right\|_2^2. \tag{4}$$

A low residual means that the node lies near the class geometry; a high residual means that the node is poorly explained by that class subspace. Unlike a black-box embedding, r_c, μ_c, B_c, and R_{ic}^{pca} can all be inspected directly.

3.4 Closed-form multi-alpha ridge boundary

PCA residuals capture class geometry, but class boundaries may still be better described by a discriminative linear separator. We therefore fit a ridge classifier in closed form. Let $Y \in \mathbb{R}^{n_{\mathrm{tr}} \times C}$ be the one-hot label matrix on the training nodes and let $\mathcal{A}$ denote the candidate regularization set. For $\alpha \in \mathcal{A}$,

$$\beta_\alpha = (F_{\mathrm{tr}} F_{\mathrm{tr}}^\top + \alpha I_{n_{\mathrm{tr}}})^{-1} Y, \qquad Z_\alpha = F F_{\mathrm{tr}}^\top \beta_\alpha, \tag{5}$$

where $\beta_\alpha \in \mathbb{R}^{n_{\mathrm{tr}} \times C}$ and $Z_\alpha \in \mathbb{R}^{n \times C}$. In the implementation, each score matrix is first normalized by a single training-split standard deviation,

$$\sigma_\alpha = \mathrm{std}\big(\{(Z_\alpha)_{ic} : i \in \mathcal{T},\ c = 1, \ldots, C\}\big), \tag{6}$$

and the residual-like ridge score is then defined by

$$R_{ic}^{\mathrm{ridge}} = -\frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \frac{(Z_\alpha)_{ic}}{\sigma_\alpha + \epsilon}. \tag{7}$$

Smaller values are therefore better, matching the PCA-residual convention.

3.5 Score fusion and prediction

Before fusion, each branch is rescaled by a training-split standard deviation:

$$\sigma_{\mathrm{pca}} = \mathrm{std}\big(\{R_{ic}^{\mathrm{pca}} : i \in \mathcal{T},\ c = 1, \ldots, C\}\big), \qquad \sigma_{\mathrm{ridge}} = \mathrm{std}\big(\{R_{ic}^{\mathrm{ridge}} : i \in \mathcal{T},\ c = 1, \ldots, C\}\big). \tag{8}$$

We then define

$$\widetilde{R}_{ic}^{\mathrm{pca}} = \frac{R_{ic}^{\mathrm{pca}}}{\sigma_{\mathrm{pca}} + \epsilon}, \qquad \widetilde{R}_{ic}^{\mathrm{ridge}} = \frac{R_{ic}^{\mathrm{ridge}}}{\sigma_{\mathrm{ridge}} + \epsilon}. \tag{9}$$

The final fused score is

$$S_{ic} = w\, \widetilde{R}_{ic}^{\mathrm{pca}} + (1 - w)\, \widetilde{R}_{ic}^{\mathrm{ridge}}, \qquad \hat{y}_i = \arg\min_c S_{ic}. \tag{10}$$

All hyperparameters, including K, the PCA dimension cap, the energy threshold, the ridge-alpha set, and w, are selected by validation accuracy. The algorithm therefore remains a validation-selected white-box classifier rather than a trained neural network.

3.6 Node-level signal atlas

The same scaffold used for prediction also produces the node-level atlas. Fisher-selected coordinates retain their dictionary-block identities, so each node can be decomposed over the explicit graph-signal dictionary.

Let $\mathcal{B}$ denote the named dictionary blocks in Eq. (1), and let $\mathcal{I}_b$ be the coordinate indices belonging to block b in the full dictionary. For each selected block

$$\mathcal{S}_b = \mathcal{S} \cap \mathcal{I}_b,$$

we define the Fisher-weighted block evidence of node i as

$$E_b(i) = \begin{cases} \dfrac{1}{|\mathcal{S}_b|} \displaystyle\sum_{j \in \mathcal{S}_b} |F_{ij}^{(0)}|\, q_j, & |\mathcal{S}_b| > 0, \\ 0, & |\mathcal{S}_b| = 0. \end{cases} \tag{11}$$

The corresponding block share is

$$\pi_b(i) = \frac{E_b(i)}{\sum_{b' \in \mathcal{B}} E_{b'}(i) + \epsilon}. \tag{12}$$

Thus, $\pi_b(i)$ is not a learned attention weight; it is a deterministic, Fisher-weighted block readout induced by the fixed graph-signal dictionary and the supervised Fisher selector.

Because the raw, low-pass, and high-pass families contain unequal numbers of blocks, family-level composition is computed by first averaging evidence within each family and then normalizing across families. For $g \in \{\text{raw}, \text{low}, \text{high}\}$, let $\mathcal{B}_g$ be the blocks in family g and define

$$\bar{E}_g(i) = \frac{1}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} E_b(i), \qquad R_g(i) = \frac{\bar{E}_g(i)}{\bar{E}_{\text{raw}}(i) + \bar{E}_{\text{low}}(i) + \bar{E}_{\text{high}}(i) + \epsilon}. \tag{13}$$

We write these three family-size-adjusted shares as $R_{\text{raw}}(i)$, $R_{\text{low}}(i)$, and $R_{\text{high}}(i)$. They are the node-level signal coordinates used in the graph-signal simplex and the dense phase portraits. Let

$$\mathcal{C}_{\mathcal{T}} = \{\, c : \exists i \in \mathcal{T} \text{ such that } y_i = c \,\}$$

denote the set of classes present in the training split.

The atlas also records the branch-wise predictions

$$\hat{y}_i^{\text{pca}} = \arg\min_c \widetilde{R}_{ic}^{\text{pca}}, \qquad \hat{y}_i^{\text{ridge}} = \arg\min_c \widetilde{R}_{ic}^{\text{ridge}}, \tag{14}$$

and the true-versus-nearest-wrong branch margins

$$M_i^{\text{pca}} = \min_{c \neq y_i} \widetilde{R}_{ic}^{\text{pca}} - \widetilde{R}_{i,y_i}^{\text{pca}}, \qquad M_i^{\text{ridge}} = \min_{c \neq y_i} \widetilde{R}_{ic}^{\text{ridge}} - \widetilde{R}_{i,y_i}^{\text{ridge}}. \tag{15}$$

If the true label y_i is absent from the training-class set $\mathcal{C}_{\mathcal{T}}$ for a particular split, these branch margins are treated as undefined and are recorded as missing values in the implementation.

Together with the final prediction $\hat{y}_i$, final correctness, degree, and the decision quadrant—both modules correct, PCA only, ridge only, or both wrong—these records form a dense node-level atlas. Aggregating them gives the dataset fingerprint used in Section 7: signal composition, PCA–Ridge decision structure, class-subspace complexity, and correct-versus-wrong signal shifts.

4 Experimental Setup

Datasets. We evaluate on six PyTorch Geometric node-classification datasets (Fey & Lenssen, 2019): Amazon-Computers, Amazon-Photo, Chameleon, Cornell, Texas, and Wisconsin. The Amazon graphs are co-purchase graphs from the benchmark studied by Shchur et al. (2018). Chameleon is loaded from `WikipediaNetwork` with `geom_gcn_preprocess=True`, and the three WebKB graphs are loaded from `WebKB`; these datasets are standard heterophily benchmarks associated with the Geom-GCN setting (Pei et al., 2020), with Chameleon originating from the Wikipedia networks of Rozemberczki et al. (2021).

Baselines. For the aligned main comparison, we rerun a disclosed candidate pool under the same dataset-repeat protocol as WG-SRC and compare against the strongest reproduced comparator for each dataset. This assignment is used only to form the comparison target; it is not used to train, tune, or select WG-SRC. The candidate pool includes GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2018), APPNP (Gasteiger et al., 2019), Correct and Smooth (Huang et al., 2021), GCNII (Chen et al., 2020), SIGN (Frasca et al., 2020), LINKX (Lim et al., 2021), and an MLP-propagation baseline. The strongest reproduced baseline is GraphSAGE for all datasets except Chameleon, where it is LINKX. All baseline epoch, model-state, and optional hyperparameter choices use validation accuracy only; the test mask is never used for training, early stopping, or selection.

Evaluation. We report mean accuracy and sample standard deviation over ten repeated class-balanced random splits. All main predictive-validity and white-box atlas figures use the six original evaluated datasets. The fingerprint-conditioned black-box probing conditions are described separately in Section 8 and Appendix G.

Diagnostic protocol. Predictor selection is separated from dataset probing. The full WG-SRC scaffold is first selected by validation accuracy and evaluated on the test set under the fixed selected configuration. The fitted scaffold is then used as a graph-dataset probe whose atlas records signal-family usage, class-subspace structure, branch behavior, margins, and error locations. Correctness-dependent atlas quantities are computed only after ordinary evaluation and are not used for training, hyperparameter selection, model selection, or black-box architecture selection.

5 Predictive Validity of the White-Box Scaffold

The atlas is meaningful only if it is produced by a classifier that captures useful predictive structure. We therefore first test whether WG-SRC remains competitive under the same repeated-split protocol as the aligned reproduced baselines.

Table 1: Mean-level companion to Figure 2. Values are mean test accuracy ± sample standard deviation over ten aligned repeats; baselines are the strongest aligned CPU reruns.

Dataset	Strongest baseline	Baseline ($n = 10$)	WG-SRC ($n = 10$)	Gain
Amazon-Computers	GraphSAGE	76.84 ± 2.28	78.71 ± 1.15	+1.87
Amazon-Photo	GraphSAGE	88.37 ± 1.86	88.76 ± 1.35	+0.39
Chameleon	LINKX	71.56 ± 1.49	72.48 ± 1.85	+0.92
Cornell	GraphSAGE	72.43 ± 6.47	75.41 ± 7.26	+2.97
Texas	GraphSAGE	84.32 ± 4.73	86.32 ± 4.08	+1.99
Wisconsin	GraphSAGE	83.33 ± 5.25	84.31 ± 4.24	+0.98
Average	–	**79.48**	**81.00**	**+1.52**

Table 1 reports the absolute repeated-split comparison, and Figure 2 shows paired split-level differences rather than only repeated means. WG-SRC achieves a positive mean gain on all six datasets and a +1.52 percentage-point average gain over the strongest aligned reproduced baseline. The split-level paired comparison shows that WG-SRC wins most paired splits on five of six datasets and ties the win count on Wisconsin; the full win-count summary is reported in Appendix A. Treating the six dataset-level gains as paired units gives a two-sided one-sample t-test $p = 0.0105$, Wilcoxon signed-rank $p = 0.0313$, and sign test $p = 0.0313$. These paired results support the use of the fitted scaffold as a diagnostic probe: the subsequent atlas is produced by a transparent model that preserves competitive predictive performance under matched repeated splits.

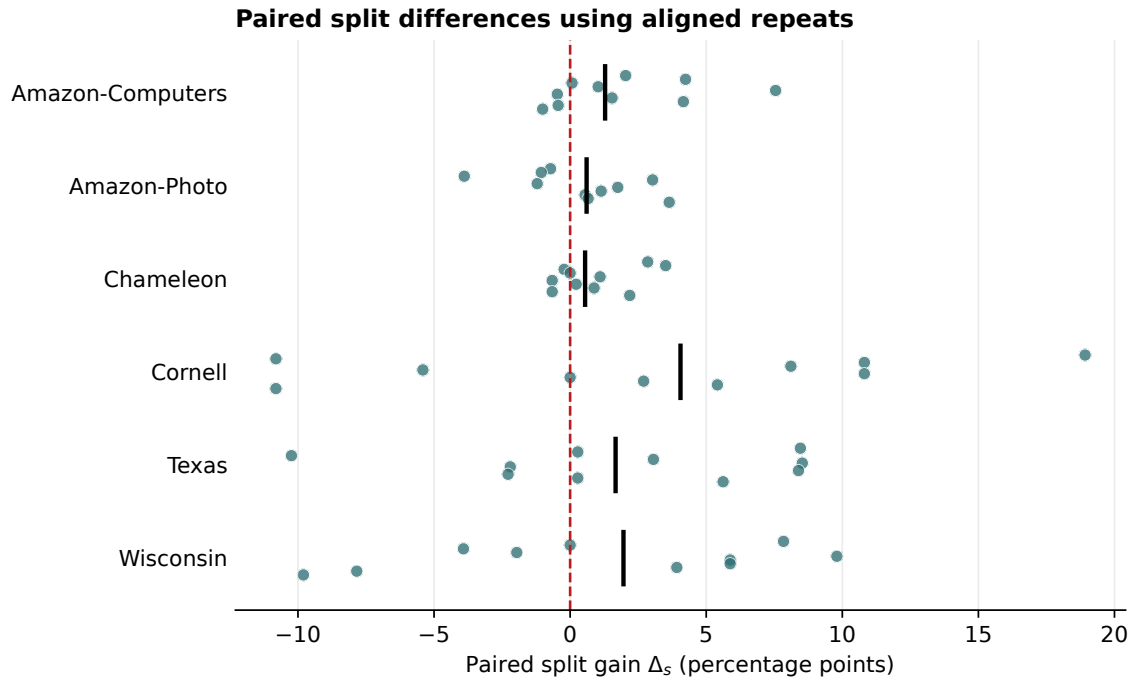


Figure 2: Split-level companion to Table 1. Each point is one matched dataset–repeat gain, $\Delta_s = \mathrm{Acc}_s^{(\mathrm{src})} - \mathrm{Acc}_s^{(\mathrm{base})}$, in percentage points; the dashed line marks zero.Summary metrics are in Appendix A.

6 Mechanistic Interventions and White-box Mechanism Checks

We evaluate simplified WG-SRC variants under the same split, seed, and train/validation/test protocol as the full scaffold. Each variant is selected by validation accuracy only and evaluated over ten matched repeats. These ablations serve as controlled interventions: by removing or isolating named signal and decision components, they test whether structural changes agree with the mechanisms measured by the atlas.

Table 2: Aligned mechanistic intervention summary. All variants are run with the same ten split/seed protocol as the final full WG-SRC scaffold experiment. Entries are mean test accuracy in percent. The bottom rows summarize cross-dataset generalization. Rank is computed within each dataset among the listed variants; lower is better. The full model is not always the best dataset-specific specialist, but it has the best average rank, the best worst-case accuracy, and is the only variant that remains in the top three on all six datasets. The simplified variants are used as diagnostic interventions, not as a redefinition of the main benchmark model.

Dataset	**Full**	Raw	No high-pass	No P^3X	No sym	PCA only	Ridge only	Best diagnostic variant
Amazon-Computers	**78.71**	63.46	79.73	76.32	73.54	76.83	78.58	No high-pass (79.73)
Amazon-Photo	**88.76**	76.88	90.27	87.16	84.78	87.15	89.12	No high-pass (90.27)
Chameleon	**72.48**	44.71	72.02	71.67	72.35	70.61	71.62	Full / near-full (72.48)
Cornell	**75.41**	77.30	71.08	75.14	74.32	66.76	77.57	Ridge only (77.57)
Texas	**86.32**	82.89	83.42	85.26	85.79	77.11	86.58	Ridge only (86.58)
Wisconsin	**84.31**	88.63	79.22	85.69	83.73	78.24	83.92	Raw only (88.63)
Mean accuracy	**81.00**	72.31	79.29	80.21	79.09	76.12	81.23	—
Worst-case accuracy	**72.48**	44.71	71.08	71.67	72.35	66.76	71.62	—
Average rank ↓	**2.33**	5.00	3.67	3.83	4.50	6.00	2.67	—
Top-3 count	**6/6**	2/6	3/6	1/6	2/6	0/6	4/6	—

Table 2 shows that no single simplified variant is uniformly reliable across datasets. Removing high-pass blocks improves the two Amazon graphs, ridge-oriented variants are strong on Cornell and Texas, and raw-only is

strongest on Wisconsin. However, these specialized gains do not transfer: raw-only collapses on Chameleon, PCA-only is weak across the board, and no simplified variant appears in the top three on every dataset.

The bottom rows of Table 2 summarize this generalist-versus-specialist pattern. The full scaffold has the best average rank, the highest worst-case accuracy, and top-three performance on all six datasets. Although Ridge-only attains a slightly higher mean accuracy, it removes the class-subspace branch needed for class-subspace complexity, PCA–Ridge complementarity, and the geometry-versus-boundary analysis in later sections. We therefore keep the full scaffold as the main diagnostic base model, while treating simplified variants as dataset-specific interventions.

These intervention results preview the atlas analysis. Amazon graphs behave as low-pass-dominated regimes in which high-pass blocks can be removed for a specialized improvement. Chameleon appears to benefit from the combined multi-hop, high-pass, PCA, and ridge structure: the full scaffold is strongest, while several near-full variants remain close. WebKB graphs are more boundary- or raw-feature sensitive, making simplified specialists useful on particular datasets. Thus, once a dataset has been measured, the same fingerprint can suggest which mechanistic simplifications or emphases are worth testing first.

7 From Node-Level Atlases to Dataset Fingerprints

The node-level atlas in Section 3.6 records signal shares, branch predictions, branch margins, correctness, degree, and PCA–Ridge decision quadrants for each evaluated node. We call the aggregation of these records an operational dataset fingerprint. It is operational because it is computed from a fixed graph-signal dictionary and fixed white-box decision modules; it is a dataset fingerprint because the same measurement procedure yields comparable signal compositions, decision geometries, class complexities, and error shifts across datasets. The fingerprint is not a claim about the data-generating process itself. It measures how the fixed WG-SRC scaffold uses a dataset.

For dataset D, let $\mathcal{U}_D$ denote the evaluation node set used for atlas reporting; in the retrospective analysis below, this is the test-node set. Using the training-class set $\mathcal{C}_\mathcal{T}$ defined in Section 3.6, write $C_\mathcal{T} = |\mathcal{C}_\mathcal{T}|$. We define the dataset-level signal means by

$$R_D = \frac{1}{|\mathcal{U}_D|} \sum_{i \in \mathcal{U}_D} R_{\text{raw}}(i), \qquad L_D = \frac{1}{|\mathcal{U}_D|} \sum_{i \in \mathcal{U}_D} R_{\text{low}}(i), \qquad H_D = \frac{1}{|\mathcal{U}_D|} \sum_{i \in \mathcal{U}_D} R_{\text{high}}(i), \tag{16}$$

the mean class-subspace complexity by

$$C_D = \frac{1}{C_\mathcal{T}} \sum_{c \in \mathcal{C}_\mathcal{T}} r_c, \tag{17}$$

and the PCA–Ridge decision fractions by

$$Q_D^{\text{ridge}} = \frac{1}{|\mathcal{U}_D|} \sum_{i \in \mathcal{U}_D} \mathbf{1}\Big\{ \hat{y}_i^{\text{ridge}} = y_i,\ \hat{y}_i^{\text{pca}} \neq y_i \Big\}, \tag{18}$$

$$Q_D^{\text{hard}} = \frac{1}{|\mathcal{U}_D|} \sum_{i \in \mathcal{U}_D} \mathbf{1}\Big\{ \hat{y}_i^{\text{ridge}} \neq y_i,\ \hat{y}_i^{\text{pca}} \neq y_i \Big\}. \tag{19}$$

Define the correct and wrong subsets of the evaluation nodes by

$$\mathcal{U}_D^{\text{correct}} = \{\, i \in \mathcal{U}_D : \hat{y}_i = y_i \,\}, \qquad \mathcal{U}_D^{\text{wrong}} = \{\, i \in \mathcal{U}_D : \hat{y}_i \neq y_i \,\}. \tag{20}$$

For the correctness-dependent high-pass shift, let

$$H_D^{\text{correct}} = \frac{1}{|\mathcal{U}_D^{\text{correct}}|} \sum_{i \in \mathcal{U}_D^{\text{correct}}} R_{\text{high}}(i), \qquad H_D^{\text{wrong}} = \frac{1}{|\mathcal{U}_D^{\text{wrong}}|} \sum_{i \in \mathcal{U}_D^{\text{wrong}}} R_{\text{high}}(i), \tag{21}$$

and define

$$\Delta H_D = H_D^{\text{wrong}} - H_D^{\text{correct}}. \tag{22}$$

Table 3: Family-size-adjusted node-level graph signal mixture. Values are average test-node signal-family shares in percent. Family composition is computed by first averaging Fisher-weighted block evidence within each signal family and then normalizing across the three families, thereby removing the trivial family-cardinality prior of the raw/low-pass/high-pass partition.

Dataset	n_{test}	Raw	Low-pass	High-pass
Amazon-Computers	13252	15.32	69.76	14.92
Amazon-Photo	7250	13.35	71.97	14.68
Chameleon	456	0.31	61.32	38.37
Cornell	37	28.12	34.99	36.89
Texas	38	29.47	30.59	39.94
Wisconsin	51	34.59	32.70	32.71

We then summarize the operational atlas of dataset D by

$$\mathbf{m}(D) = \left[R_D,\, L_D,\, H_D,\, C_D,\, Q_D^{\mathrm{ridge}},\, Q_D^{\mathrm{hard}},\, \Delta H_D\right]. \tag{23}$$

In the present paper, these quantities are reported on test nodes to characterize how the fixed white-box probe behaves on the dataset after standard supervised evaluation. The fingerprint contains two kinds of atlas quantities. The signal-composition and class-geometry terms R_D, L_D, H_D, and C_D are correctness-free summaries of how the fitted scaffold uses the evaluated dataset. By contrast, Q_D^{ridge}, Q_D^{hard}, and ΔH_D are correctness-dependent audit quantities: they locate where the already evaluated predictor succeeds or fails in the same atlas coordinates. These quantities are used only as retrospective same-dataset diagnostic evidence, not as inputs to training, hyperparameter selection, blind model choice, or test-set-tuned redesign. If related diagnostic procedures are used prospectively, correctness-dependent quantities should be computed only on training or validation nodes.

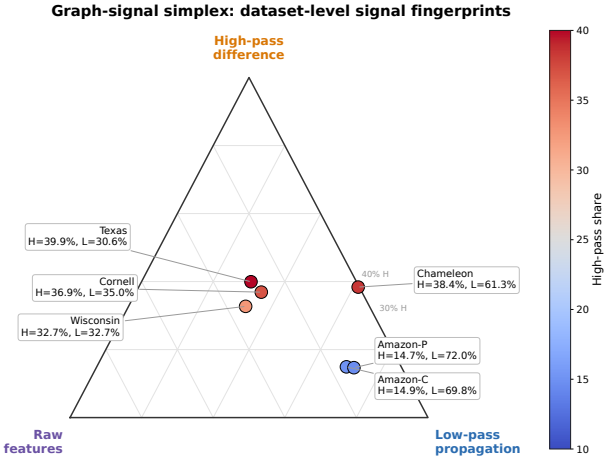
7.1 Graph-signal fingerprints

Table 3 and Figure 3a report family-size-adjusted signal-family shares, so the reported composition is not mechanically driven by the unequal 1:5:3 block counts of the raw/low-pass/high-pass partition. Amazon-Computers and Amazon-Photo remain low-pass dominated, but at a corrected level of roughly 70–72% rather than the unadjusted 84–85%. Chameleon shifts toward a mixed low-pass/high-pass profile, while Cornell, Texas, and Wisconsin move into a more raw- and high-pass-balanced regime. Thus, the classical homophily/heterophily dichotomy can be refined into a measurable signal composition: a graph may still be low-pass dominated, yet retain a nontrivial raw or high-pass component after correcting for family cardinality.

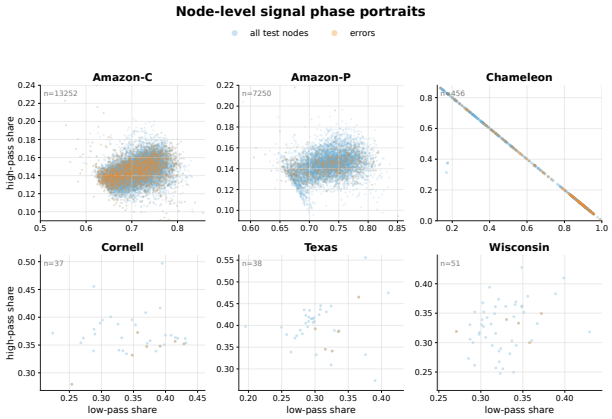
Figures 3b–3d provide the node- and class-level evidence behind the dataset fingerprint. Figure 3b localizes errors inside the low-pass/high-pass plane rather than only reporting dataset-level averages: Amazon errors concentrate away from the dominant low-pass core, whereas Chameleon occupies a broader mixed regime. Figure 3c separates geometry-driven, boundary-driven, easy, and hard nodes, showing why ridge-style boundary correction is relevant on the WebKB graphs. Figure 3d shows that Chameleon requires much larger class subspaces than Amazon and WebKB, confirming that its selected graph-signal geometry is substantially more complex. Together, these panels close the main-text evidence chain from signal composition to node-level error structure, decision-phase behavior, and class-subspace complexity.

8 Fingerprint-Guided Diagnostic Guidance

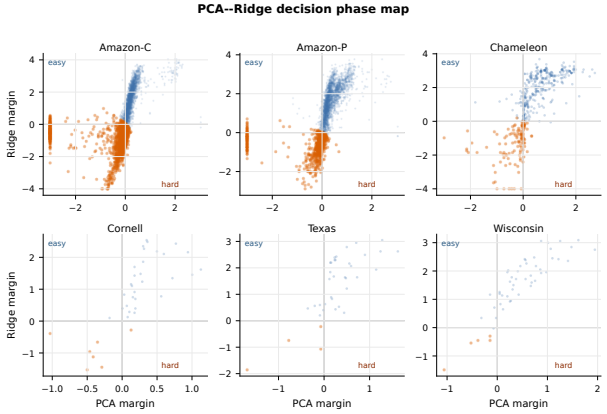
The mechanistic atlas turns a dataset from a single accuracy number into an operational diagnostic object. Given the fingerprint $\mathbf{m}(D)$ in Eq. (23), WG-SRC asks which signal families, class-subspace structures, and decision mechanisms are associated with success or failure under the fitted white-box scaffold. The claim is not automatic architecture search: the atlas provides post-evaluation, same-dataset mechanism hypotheses that must be checked by aligned interventions and validation-based model development. The operational



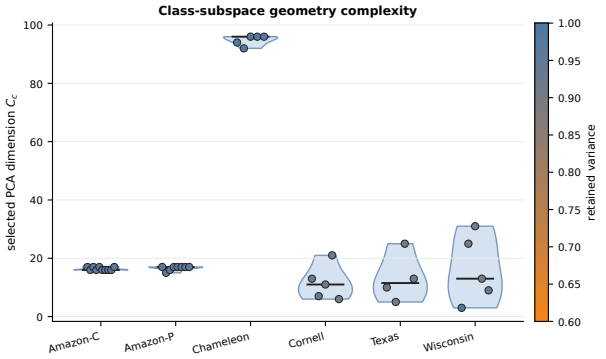
(a) Dataset-level signal simplex.



(b) Node-level signal phase portraits.



(c) PCA–Ridge decision phase map.



(d) Class-subspace geometry complexity.

Figure 3: Main-text atlas evidence for operational dataset fingerprints. (a) The signal simplex summarizes family-size-adjusted raw, low-pass, and high-pass signal composition. (b) Node-level phase portraits localize errors inside the low-pass/high-pass plane. (c) PCA–Ridge phase maps separate geometry-driven, boundary-driven, easy, and hard nodes. (d) Class-subspace complexity shows the selected PCA dimension needed by each class. Together, the four panels connect dataset-level signal composition, node-level error structure, decision-branch behavior, and class-subspace geometry.

claim is that the atlas can prioritize follow-up analysis after evaluation: it identifies which mechanisms should be suppressed, preserved, strengthened, or further tested first on the same measured dataset. These mechanism hypotheses are then checked by aligned interventions and by subsequent validation-based model development.

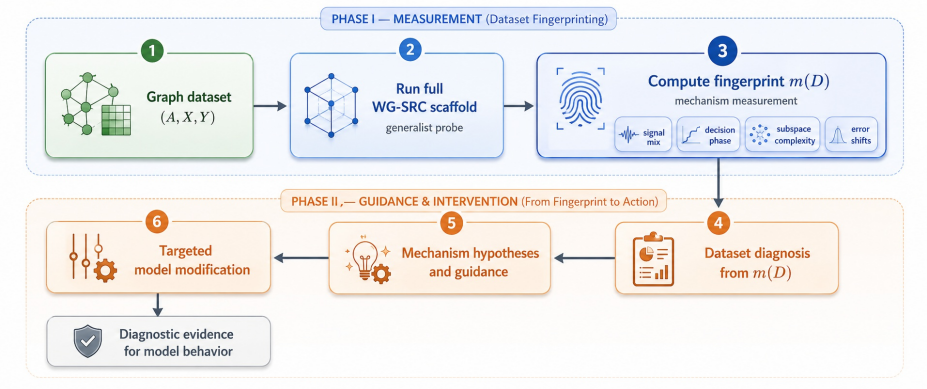


Figure 4: Mechanism atlas as a dataset diagnostic probe. WG-SRC runs the full white-box scaffold, computes $\mathbf{m}(D)$, and uses it to generate post-evaluation, same-dataset mechanism hypotheses checked by aligned interventions and fingerprint-conditioned black-box probing.

Table 4: From operational dataset fingerprint to same-dataset diagnostic guidance. Each row maps an atlas signature to a mechanism diagnosis and a first dataset-specific modification direction. The table is an operational consistency map: the proposed directions are checked against aligned white-box interventions and fingerprint-conditioned black-box component probes

Atlas signature	Dataset diagnosis	Diagnostic guidance
High L_D, low H_D, and positive ΔH_D	High-pass components may act as noise under the fitted scaffold	Remove, downweight, gate, or regularize high-pass blocks
Near-zero R_D and high C_D	Raw-block evidence is weak and class geometry is complex	Preserve graph-derived signals and improve class-specific subspace modeling
High H_D and negative ΔH_D	Correct nodes show stronger high-pass evidence	Preserve or adaptively gate high-pass differences rather than removing them globally
High Q_D^{ridge}	The ridge branch corrects some PCA-branch failures	Strengthen the discriminative head or improve PCA–Ridge boundary fusion
High Q_D^{hard}	Both geometry and boundary branches fail on many nodes	Investigate new signal blocks, hard-node treatment, uncertainty handling, or graph rewiring
High R_D and low-to-moderate C_D	Raw-block evidence is strong under the fitted scaffold	Use raw-preserving skip design, weak propagation, or propagation gating

Table 4 summarizes the diagnostic mapping from atlas signatures to first modification directions. The supporting white-box evidence comes from the aligned interventions in Table 2 and the correct-versus-wrong signal shifts in Appendix B. When the atlas suggests that a mechanism is operationally harmful or necessary under the fitted scaffold, the corresponding fixed intervention usually moves performance in the predicted direction: Amazon benefits from removing high-pass blocks, Chameleon fails under raw-only simplification, Texas shows strong boundary-oriented behavior, and Wisconsin matches the raw-feature diagnosis. Thus, the fingerprint is not only descriptive; it identifies which mechanisms are worth testing first for a measured dataset.

Two measured conditions in the compact black-box summary are processed prototypes used only for component probing. Amazon-Photo-LP+ strengthens a low-pass regime by label-free mutual cosine-kNN edge densification, and Wisconsin-R+ creates a raw-enhanced WebKB stress condition by label-free rewiring/dropout processing. They are not used in the main predictive-validity claim; they are controlled stress tests for asking whether sharper measured dataset conditions expose corresponding black-box component biases. Construction details are given in Appendix G.

Table 5 gives the corresponding minimum black-box evidence. The full family-by-family component tables are reported in Appendix C, but the key pattern is already visible in the compact summary: different measured fingerprints repeatedly favor different architectural biases under fixed black-box component probes.

Table 5: Main-text compact summary of fingerprint-conditioned black-box component probing. The table reports qualitative tendencies under fixed component ablations across LINKX, GNNMoE, SGFormer, GOAT, NodeFormer, and tuned GraphSAGE. Its purpose is to test whether measured dataset fingerprints organize black-box component behavior, rather than to rank model families. Full tables in Appendix C.

Measured condition	Fingerprint diagnosis	Favored tendency	Usually disfavored tendency
Amazon-Photo-LP+	Low-pass-dominant stress-test regime	Shallow local propagation; original-graph or graph-branch behavior; raw-preserving or no-high-pass designs	Pure global attention; latent-only rewiring; unnecessary high-pass complexity
Chameleon	Mixed low/high-pass, near-zero raw signal, high class-geometry complexity	Graph-local structure; graph-branch dominance; selective structural bias	Token-only, global-only, latent-only, or raw-dominant variants
Wisconsin-R+	Raw-enhanced WebKB regime with boundary sensitivity	Root/raw-preserving shallow models; light original-graph use	No-root ablations; deep structural complexity
Texas	Boundary-sensitive and high-pass-balanced WebKB regime	Shallow message passing; stronger discriminative or global mixing	Over-fused variants; extra depth that dilutes the discriminative branch
Cornell	Balanced WebKB control case	Simple original-graph bias; shallow propagation; light or no fusion	Deep heads, no-root ablations, or overly complex fusion modules

Thus, the black-box probing evidence is not used to rank modern graph models. Its role is narrower and diagnostic: once WG-SRC assigns an operational fingerprint to a measured dataset, fixed component ablations show which architectural directions are worth modifying first. This closes the main-text evidence chain from white-box fingerprint, to aligned intervention, to fingerprint-conditioned black-box behavior.

9 Limitations

WG-SRC trades speed for an explicit audit trail: graph-signal construction, Fisher selection, class-wise PCA, ridge solves, and validation search all add cost, so Appendix D reports a computational profile rather than a speed claim. The method is modular: specialized variants can outperform the full scaffold on some datasets, while the full model is kept as the analytically complete atlas probe. The baseline comparison is limited to the disclosed aligned reproduced suite, and the black-box probing study is an initial same-dataset demonstration; final model development still requires validation under the target dataset protocol.

10 Conclusion

We presented WG-SRC, a white-box signal-subspace scaffold for graph node classification and dataset diagnosis. WG-SRC replaces hidden message passing with explicit raw, low-pass, and high-pass graph-signal blocks, Fisher selection, class-wise PCA subspaces, and a closed-form ridge boundary, so the same fitted variables used for prediction also define an operational dataset-under-scaffold fingerprint.

Across six datasets under aligned repeated splits, WG-SRC remains competitive with selected reproduced baselines while preserving the structure needed for diagnosis. Its fingerprints separate low-pass-dominated Amazon graphs, mixed high-pass and class-geometrically complex Chameleon behavior, and raw- or boundary-sensitive WebKB graphs. Aligned interventions support these measurements: high-pass removal helps when high-pass evidence is noise-like, graph-derived signals matter when raw features are insufficient, and ridge-oriented decisions matter when class geometry alone is inadequate.

The contribution is therefore a reproducible diagnostic scaffold rather than a leaderboard or causal claim. After validation-selected evaluation, WG-SRC provides mechanism hypotheses for same-dataset analysis and targeted model modification, and the black-box component probes in Appendix C show how the measured fingerprints can organize architectural-bias tests beyond the white-box model.

References

K. H. R. Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. ReduNet: A white-box deep network from the principle of maximizing rate reduction. *Journal of Machine Learning Research*, 23(114): 1–103, 2022.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735, 2020.

Xuan Chen, Jiayu Zhou, Siyuan Yu, and Qi Xuan. Mixture of experts meets decoupled message passing: Towards general and adaptive node classification. arXiv preprint arXiv:2412.08193, 2024.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations*, 2021.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Benjamin Chamberlain, Michael M. Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks. arXiv preprint arXiv:2004.11198, 2020.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pp. 1024–1034, 2017.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.

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

Kezhi Kong, Jiani Chen, John Kirchenbauer, Ruiqi Ni, C. Bayan Bruss, and Tom Goldstein. GOAT: A global transformer on large-scale graphs. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 17375–17390, 2023.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Liu Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Advances in Neural Information Processing Systems*, 2021.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11): 559–572, 1901.

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

Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.

Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. In *Relational Representation Learning Workshop, NeurIPS*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Peng Wang, Huan Liu, Dinesh Pai, Yaodong Yu, Zhihui Zhu, Qing Qu, and Yi Ma. A global geometric analysis of maximal coding rate reduction. In *International Conference on Machine Learning*, 2024.

Qitian Wu, Wentao Zhao, Zenan Li, David Wipf, and Junchi Yan. NodeFormer: A scalable graph structure learning transformer for node classification. In *Advances in Neural Information Processing Systems*, 2022.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haonan Jiang, Yatao Bian, and Junchi Yan. SGFormer: Simplifying and empowering transformers for large-graph representations. In *Advances in Neural Information Processing Systems*, 2023.

Yaodong Yu, K. H. R. Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, 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. In *Advances in Neural Information Processing Systems*, 2020.

Reproducibility Statement

The package includes the LaTeX source, filtered summary tables, figures, CSV files used to generate the tables and atlas figures, and the black-box component-probing outputs used in the fingerprint-conditioned analyses. All predictive-validity tables and white-box atlas figures use the six original evaluated datasets. The fingerprint-conditioned black-box probing summaries additionally include the processed prototype conditions described in Appendix G; these prototypes are not used in the main predictive-validity claim. Model and baseline selections are validation-only within each run, and paired split stability is computed by matching dataset and repeat index. An anonymized supplementary source package contains the implementation scripts, configuration files, split indices, random seeds, processed experiment outputs, and figure-generation files needed to reproduce the reported tables and figures.

Broader Impact Statement

This work aims to make graph learning more transparent. Potential positive impacts include easier auditing of graph models and better diagnosis of when graph propagation is harmful. Potential negative impacts are similar to those of node-classification systems generally: if applied to sensitive social, financial, or biological networks without care, predictions and explanations could still be misused. The proposed atlas should be treated as a diagnostic aid, not as a guarantee of fairness or causality.

A Paired Random-Split Stability

The main text reports absolute repeated-split accuracy in Table 1 and visualizes split-level paired differences in Figure 2. This appendix reports the paired win-count table and keeps the formal paired-effect definitions and significance-testing details.

Table 6: Paired random-split stability. For split s, $\Delta_s = Acc_{\mathrm{SRC},s} - Acc_{\mathrm{base},s}$ in percentage points, paired by dataset and repeat index. Baseline values are taken from the aligned CPU baseline rerun.

Dataset	WG-SRC wins	Mean Δ	Median Δ	Range Δ
Amazon-Computers	7/10	+1.87	+1.28	[-1.01, +7.55]
Amazon-Photo	6/10	+0.39	+0.61	[-3.89, +3.64]
Chameleon	7/10	+0.92	+0.55	[-0.66, +3.51]
Cornell	7/10	+2.97	+4.05	[-10.81, +18.92]
Texas	7/10	+1.99	+1.67	[-10.24, +8.53]
Wisconsin	5/10	+0.98	+1.96	[-9.80, +9.80]

A.1 Paired Significance Testing

Because WG-SRC and the strongest baseline are evaluated on matched dataset–repeat pairs, the appropriate comparison is paired. For dataset D and repeat s, we define the paired gain as

$$d_{D,s} = 100 \left(\mathrm{Acc}_{D,s}^{(\mathrm{src})} - \mathrm{Acc}_{D,s}^{(\mathrm{base})} \right). \tag{24}$$

where gains are measured in percentage points. Table 7 summarizes the paired effect size on each dataset without cluttering the main text with six separate per-dataset hypothesis tests.

All six dataset-level mean gains are positive. To avoid treating the 60 split-level differences as fully independent observations, we report the six dataset-level mean gains as the paired units for the aggregate stability summary. A two-sided one-sample t-test over these six dataset-level gains gives $p = 0.0105$, a two-sided Wilcoxon signed-rank test gives $p = 0.0313$, and a two-sided sign test for all six gains being positive gives $p = 0.0313$. We use these tests as compact paired-protocol evidence rather than as a standalone independent-dataset benchmark claim. Thus, the aggregate paired evidence is consistent with positive dataset-level gains under the aligned repeated-split protocol and supports the predictive validity of the fitted scaffold as a diagnostic probe.

B Additional Atlas Views

The signal simplex, dense node-level signal phase portraits, PCA–Ridge decision phase map, and class-subspace geometry complexity plot are included in the main text as Figure 3. This appendix retains only the additional correct-versus-wrong signal-shift view, which provides a complementary retrospective error diagnostic.

B.1 Correct-vs-wrong signal shifts

Table 7: Paired effect summary against the strongest aligned baseline. Gains are reported in percentage points. The effect size d_z is paired Cohen's d, computed as the mean paired gain divided by the standard deviation of the split-level paired gains. The final row treats the six dataset-level mean gains as the paired units.

Dataset	Mean gain	95% CI	d_z
Amazon-Computers	+1.87	[−0.07, +3.81]	0.69
Amazon-Photo	+0.39	[−1.20, +1.98]	0.18
Chameleon	+0.92	[−0.14, +1.98]	0.62
Cornell	+2.97	[−4.05, +10.00]	0.30
Texas	+1.99	[−2.35, +6.33]	0.33
Wisconsin	+0.98	[−3.84, +5.81]	0.15
Dataset-level aggregate	+1.52	[+0.54, +2.50]	1.62

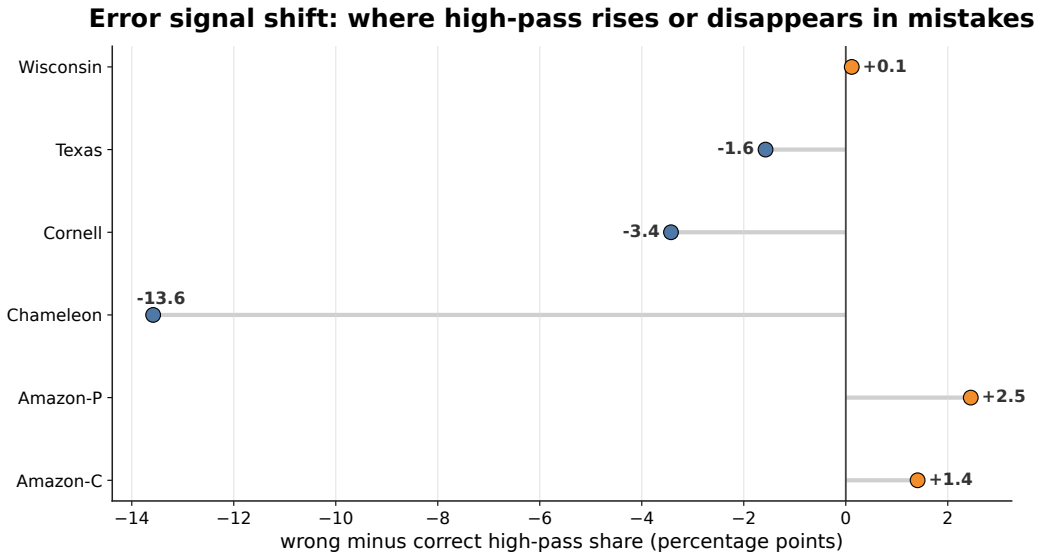


Figure 5: Error signal shift. The lollipop value is the family-size-adjusted high-pass share of wrong nodes minus the family-size-adjusted high-pass share of correct nodes. Positive values mean errors have more high-pass mass; negative values mean errors lose high-pass mass relative to correct nodes.

Figure 5 shows that errors are not uniformly distributed in signal space. On Amazon graphs, errors carry more high-pass mass than correct nodes, consistent with errors being atypical relative to the dominant low-pass mechanism. On Chameleon, the shift goes in the opposite direction: correct nodes rely more on high-pass differences, while errors are relatively more low-pass. This directly supports the design choice of including $X - P_{\text{row}}X$ and $P_{\text{row}}X - P_{\text{row}}^2X$ in the dictionary.

C Fingerprint-conditioned black-box component probing

The main text uses fingerprint-conditioned black-box probing as external evidence that the measured dataset condition organizes architectural behavior beyond the WG-SRC scaffold. This appendix provides the detailed component-level evidence behind that claim. The goal is not to rank black-box graph models or to perform automatic architecture search. Instead, after WG-SRC has assigned a fingerprint to a measured dataset, we use fixed within-family component ablations to ask which architectural biases are favored under that same dataset condition.

The probing set contains three original reference datasets (Chameleon, Texas, and Cornell) and two label-free processed prototypes. Amazon-Photo-LP+ strengthens a low-pass regime by mutual cosine-kNN

edge densification, and WISCONSIN-R+ creates a raw-enhanced WebKB stress condition by rewiring/dropout processing. Construction details and scope restrictions are reported in Appendix G.

Table 8 summarizes the qualitative fingerprint-conditioned tendencies before the family-by-family component tables. It should be read as a compact guide to the detailed ablations in Tables 9 and 10, not as a prospective model-selection rule.

Table 8: Compact fingerprint-conditioned summary of black-box component tendencies. This is a post-evaluation same-dataset summary, not a prospective model-selection rule or an architecture-search table.

Dataset	Fingerprint type	Favored bias	Usually disfavored
Amazon-Photo-LP+	Low-pass-dominant stress test	Shallow local propagation; original-graph / graph-branch behavior; simple raw-preserving or no-high-pass designs	Pure global attention; latent-only rewiring; deep heads; unnecessary high-pass complexity
Chameleon	Mixed low/high-pass; near-zero raw; high class-geometry complexity	Graph-local structure; graph-branch dominance; no-root aggregation; selective structural bias	Token-only or global-only variants; latent-only rewiring; raw-dominant bias
Wisconsin-R+	Raw-enhanced WebKB regime	Root/raw-preserving shallow models; one-layer message passing; light original-graph use	No-root ablations; deep structural complexity; unnecessary latent/global fusion
Texas	Boundary-sensitive, high-pass-balanced WebKB regime	Shallow message passing; stronger discriminative/global mixing; original-graph branch	Over-fused full variants; extra depth when it dilutes the discriminative branch
Cornell	Balanced WebKB control case	Simple original-graph bias; shallow propagation; light or no fusion	Deep heads; no-root ablations; overly complex global/fusion modules

What is being tested. The LINKX, GNNMoE, SGFormer, GOAT, NodeFormer, and tuned GraphSAGE studies all ask the same question: once a dataset fingerprint has been measured, which kind of black-box architectural bias is favored by that measured condition? We fix the repeated-split protocol and use fixed component ablations within each model family. The resulting tables should therefore be read as a fingerprint-conditioned probing map for same-dataset model analysis rather than as a universal ranking of black-box architectures.

For compactness, we abbreviate variant names inside the tables: nf-linear = no_fusion_linear, no-hp-exp = no_hp_expert, g-only = graph_only, tok-only = token_only, glob-only = global_only, loc-only = local_only, s-head = shallow_head, n-root = no_root, and 1-layer = one_layer.

Table 9: Fingerprint-conditioned black-box probing organized by model family (part I). Each family entry reports the strongest variant and one informative contrast on that measured dataset.

Dataset / role	Fingerprint type	Fingerprint summary	LINKX	GNNMoE	SGFormer
Amazon-Photo-LP+	Low-pass-dominant processed prototype	L dominant; high-pass error-associated	**top:** nf-linear (0.766) **vs.:** full (0.229)	**top:** no-hp-exp (0.698) **vs.:** full (0.261)	**top:** g-only (0.825) **vs.:** tok-only (0.210)
Chameleon	Mixed graph-structured heterophilic regime	Near-zero raw share; substantial high-pass; high class complexity	**top:** struct-only (0.636) **vs.:** x-only (0.468)	**top:** hard-top1 (0.598) **vs.:** full (0.555)	**top:** g-only (0.708) **vs.:** tok-only (0.223)
Wisconsin-R+	Raw-enhanced processed prototype	Strong raw signal; some boundary sensitivity	**top:** nf-linear (0.842) **tie:** x-only (0.842)	**top:** no-hp-exp (0.834) **vs.:** hard-top1 (0.697)	**top:** tok-only (0.739) **vs.:** g-only (0.456)
Texas	High-pass- and boundary-sensitive WebKB	High-pass active; clearer boundary correction	**top:** nf-linear (0.849) **vs.:** full (0.765)	**top:** no-hp-exp (0.868) **vs.:** full (0.838)	**top:** tok-only (0.703) **vs.:** g-only (0.565)
Cornell	Balanced WebKB control case	More balanced raw/high-pass; less extreme than Texas	**top:** x-only (0.773) **vs.:** full (0.735)	**top:** no-hp-exp (0.768) **tie:** full (0.768)	**top:** tok-only (0.592) **vs.:** g-only (0.397)

Table 10: Fingerprint-conditioned black-box probing organized by model family (part II). Each family entry reports the strongest variant and one informative contrast on that measured dataset.

Dataset / role	Fingerprint type	Fingerprint summary	GOAT	NodeFormer	GraphSAGE
Amazon-Photo-LP+	Low-pass-dominant processed prototype	L dominant; high-pass error-associated	**top:** loc-only (0.261) **note:** family weak overall	**top:** g-only (0.757) **vs.:** latent-only (0.158)	**top:** s-head (0.896) **vs.:** n-root (0.882)
Chameleon	Mixed graph-structured heterophilic regime	Near-zero raw share; substantial high-pass; high class complexity	**top:** full (0.504) **vs.:** glob-only (0.223)	**top:** full (0.469) **vs.:** latent-only (0.229)	**top:** n-root (0.684) **vs.:** 1-layer (0.613)
Wisconsin-R+	Raw-enhanced processed prototype	Strong raw signal; some boundary sensitivity	**top:** glob-only (0.754) **vs.:** full (0.753)	**top:** g-only (0.828) **vs.:** latent-only (0.682)	**top:** 1-layer (0.822) **vs.:** n-root (0.616)
Texas	High-pass- and boundary-sensitive WebKB	High-pass active; clearer boundary correction	**top:** glob-only (0.695) **vs.:** loc-only (0.624)	**top:** nf-linear (0.830) **vs.:** full (0.695)	**top:** 1-layer (0.849) **note:** strongest overall
Cornell	Balanced WebKB control case	More balanced raw/high-pass; less extreme than Texas	**top:** s-head (0.592) **vs.:** full (0.514)	**top:** nf-linear (0.765) **vs.:** g-only (0.759)	**top:** 1-layer (0.746) **vs.:** n-root (0.573)

D Computational Profile

Table 11 reports a computational profile for completeness. It is not used as a main-text speed claim. Baseline runtime and accuracy are taken from the aligned CPU baseline rerun; WG-SRC accuracy is standardized to the ten-repeat main-table value. The purpose of this table is to document the cost of the white-box audit trail: explicit graph-signal construction, Fisher selection, class-wise PCA fitting, ridge solves, and validation search.

Table 11: Appendix computational profile. Baseline runtime and accuracy are taken from the aligned CPU baseline rerun; WG-SRC accuracy is standardized to the ten-repeat main-table value. This table documents runtime cost and is not intended as a speedup claim.

Dataset	Best baseline	Baseline time(s)	WG-SRC time(s)	Ratio	WG-SRC acc.
Amazon-Computers	GraphSAGE	208.5	292.9	1.4×	78.71
Amazon-Photo	GraphSAGE	81.3	136.4	1.7×	88.76
Chameleon	LINKX	3.6	62.1	17.5×	72.48
Cornell	GraphSAGE	0.8	4.5	5.7×	75.41
Texas	GraphSAGE	0.8	4.1	5.2×	86.32
Wisconsin	GraphSAGE	0.9	6.0	6.8×	84.31

E Implementation Details

The Fisher coordinate budget was selected from $K \in \{4000, 5000, 6000, 8000\}$. For a dataset with dictionary dimension $p = 9d$, a candidate budget larger than p was clipped to

$$K_{\text{eff}} = \min(K, p),$$

so Fisher coordinate selection retained the top K_{eff} coordinates and retained all dictionary coordinates when $K \geq p$. The selected coordinate set $\mathcal{S}$ in Eq. (2) therefore has size $|\mathcal{S}| = K_{\text{eff}}$ in this corner case. The PCA maximum dimension was selected from $r_{max} \in \{32, 48, 64, 96\}$, with energy threshold $\eta \in \{0.90, 0.95, 0.99\}$. Ridge regularizer sets were selected from $\{0.01, 0.1, 1.0\}$, $\{0.05, 0.5, 5.0\}$, and $\{0.1, 1.0, 10.0\}$. The fusion weight was selected from $w \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. All selections were made using validation accuracy. The corresponding implementation and experiment scripts are included in the anonymized supplementary package described in the Reproducibility Statement.

F Class-Pair Geometry and Confusion

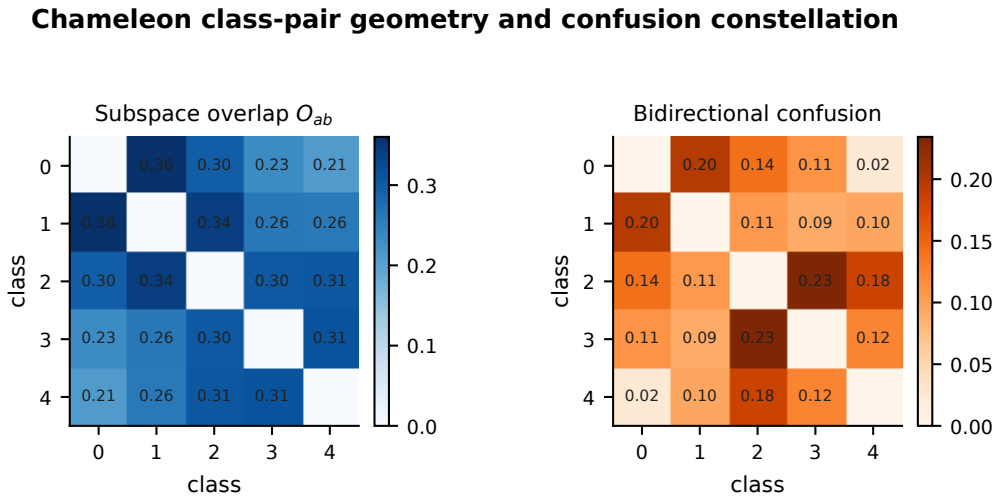


Figure 6: Chameleon class-pair geometry and confusion constellation. The left panel shows pairwise PCA subspace overlap; the right panel shows bidirectional confusion. This appendix figure illustrates how the same subspace objects used for prediction can also diagnose class-pair errors.

G Processed Prototype Construction and Scope

Table 12 documents the processed prototype datasets used only in the fingerprint-conditioned black-box probing analysis. These datasets are not part of the main predictive-validity claim and are not introduced as new benchmark datasets. Their role is only to provide label-free controlled stress-test environments under which black-box component behavior can be compared under clearer measured dataset conditions.

Table 12: Construction details of the processed prototype datasets used in the black-box probing analysis. Edge counts are undirected edge counts after processing.

Dataset	Construction	Key parameters	Processed size	Scope
Amazon-Photo-LP+	Amazon-Photo plus mutual cosine-kNN edge densification in feature space	$k = 10$; label-free	7650 nodes, 119250 undirected edges	Used only as a low-pass stress-test prototype in black-box probing
Wisconsin-R+	Wisconsin plus degree-preserving rewiring; if rewiring fails, use uniform edge-dropout fallback	rewire fraction = 0.20; fallback dropout fraction = 0.15; label-free	249 nodes, 449 undirected edges	Used only as a raw-enhanced stress-test prototype in black-box probing
Chameleon	Original processed dataset retained as reference	none	2277 nodes, 31396 undirected edges	Reference dataset in black-box probing
Texas	Original processed dataset retained as reference	none	183 nodes, 287 undirected edges	Reference dataset in black-box probing
Cornell	Original processed dataset retained as reference	none	183 nodes, 278 undirected edges	Reference dataset in black-box probing