NODE SIGNAL ATLAS: MAPPING PREDICTIVE SIGNAL BEYOND HOMOPHILY

Anonymous authors
Paper under double-blind review

ABSTRACT

Node classification benchmarks often favor very different model interfaces, yet these differences are commonly summarized by homophily. We introduce the *Node Signal Atlas* (NSA), a lightweight pre-training diagnostic that characterizes a dataset through three marginal signal channels: the *Feature Signal Score* (FSS), the *Topology Signal Score* (TSS), and the *Label Smoothness Score* (LSS). Each coordinate is computed using the same fixed logistic-regression probe on inexpensive annulus-based descriptors, without training a GNN.

On a controlled 24-dataset synthetic suite, NSA responds systematically to feature, structural-role, and local label-context manipulations, captures mixed-signal regimes, and exposes an interaction-only boundary outside its marginal scope. Across 20 real benchmarks, NSA separates datasets with nearly identical homophily but substantially different signal profiles and model behavior. Finally, on a frozen cross-generator architecture-selection benchmark, NSA improves over scalar node homophily, while a richer homophily profile is a stronger standalone baseline; combining the two yields the best low-budget triage. NSA thus complements homophily by describing *where predictive signal is accessible*.

1 INTRODUCTION

Node classification benchmarks often produce surprising outcomes: simple feature-only MLPs can match or outperform sophisticated message-passing architectures (Lim et al., 2021; Luan et al., 2024a), diffusion methods such as label propagation or APPNP (Gasteiger et al., 2019) can be unexpectedly strong, and graph-aware models help only on some datasets (Zhu et al., 2020). These differences are usually discussed through homophily, the tendency of adjacent nodes to share labels. Homophily is informative, but incomplete: datasets with similar homophily can expose very different sources of predictive signal and consequently present very different learning problems (Ma et al., 2022; Luan et al., 2024a; Platonov et al., 2023b).

This paper takes a signal-centered view of node classification. Rather than asking first which architecture wins, we ask: *what predictive signal does the dataset expose before any graph model is trained?* We focus on three broad marginal channels aligned with common model interfaces: feature signal, accessible without graph propagation; unlabeled structural-role signal, carried by graph topology itself; and local label-context signal, describing how informative neighboring labels are about a target node. These channels are not intended to exhaust all possible predictive structure: richer models may exploit interactions among them that no individual marginal probe detects. Instead, they separate three common sources of signal that a single label–adjacency statistic cannot. Strong feature signal can support competitive feature-only models, structural signal can make topology-aware processing useful, and strong local label context can favor propagation-based methods.

We introduce the *Node Signal Atlas (NSA)*, a lightweight pre-training diagnostic that embeds a node-classification problem into a three-dimensional signal space. NSA assigns each dataset a *Feature Signal Score* (FSS), a *Topology Signal Score* (TSS), and a *Label Smoothness Score* (LSS), measuring feature accessibility, unlabeled structural-role information, and local label-context information, respectively. Each coordinate is computed using the same fixed logistic-regression probe applied to inexpensive annulus-based descriptors under a uniform protocol across datasets. Probe performance is measured by chance-normalized macro-accuracy, placing datasets with different numbers of classes and class imbalance on a common scale. NSA trains no GNN and is computed once, before downstream architecture training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

NSA is not intended to replace homophily, validation, or model search. Its primary purpose is to make the learning problem itself more interpretable. On a controlled 24-dataset synthetic suite, the three coordinates respond systematically to feature, topology, and local label-context manipulations, behave continuously under mixed signal, and expose a deliberate interaction-only boundary case outside NSA's three marginal channels. On 20 real benchmarks, the atlas separates datasets with nearly identical homophily but substantially different signal profiles and model behavior. In particular, homophily is most closely aligned with local label context, while feature accessibility and structural-role information vary substantially even among graphs with similar homophily.

As a secondary operational test, we ask whether these descriptors can help prioritize architectures under a limited evaluation budget. On a separate frozen cross-generator architecture-selection benchmark, NSA substantially improves over scalar node homophily, while a richer four-statistic homophily profile is a stronger standalone selector. Crucially, the two views are complementary: combining NSA with homophily improves architecture triage beyond either representation alone. NSA is therefore best viewed not as a replacement for existing graph descriptors, but as an additional coordinate system describing *where predictive signal is accessible*.

Contributions.

- **Node Signal Atlas.** We introduce a lightweight pre-training characterization of node-classification problems through three interpretable coordinates—FSS, TSS, and LSS—measuring feature accessibility, unlabeled structural-role signal, and local label-context signal using fixed probes without training a GNN.
- **Controlled and real-world validation beyond homophily.** A 24-dataset synthetic suite shows that NSA responds systematically to controlled signal strength, represents mixed-signal problems continuously, and exposes an interaction-only boundary outside its marginal scope. Across 20 real benchmarks, NSA distinguishes datasets with nearly identical homophily but substantially different signal composition and downstream behavior.
- **Complementary architecture triage.** As a secondary operational validation, a frozen cross-generator benchmark shows that NSA substantially improves over scalar node homophily, while a richer four-statistic homophily profile is a stronger standalone baseline. Combining the two descriptions further improves low-budget architecture triage, showing that they capture complementary aspects of the learning problem.

2 Background

Why Homophily Is Not Enough. Homophily, the tendency of adjacent nodes to share labels, has long been used to characterize node-classification datasets and to motivate message passing (Kipf & Welling, 2017; Hamilton et al., 2017). Many refinements have been proposed, including node homophily, class homophily, adjusted homophily, and label informativeness (Zhu et al., 2020; Pei et al., 2019; Lim et al., 2021; Platonov et al., 2023a). These quantities are useful summaries of label agreement across edges, but they remain limited as dataset diagnostics. In particular, they are most closely tied to local label smoothness, and therefore primarily reflect the signal exploited by diffusion-style methods. They do not directly measure whether node features are discriminative or whether unlabeled structural roles correlate with class labels.

This limitation is visible empirically. Datasets with similar homophily can exhibit sharply different model behavior: on some, feature-only MLPs remain competitive; on others, propagation methods dominate; and on others, graph-aware models benefit from structural information (Platonov et al., 2023b; Luan et al., 2024a; Ma et al., 2022). Recent work has therefore argued for richer notions of graph characterization, including label informativeness, node distinguishability, classifier-based aggregation diagnostics, and multi-aspect homophily measures (Platonov et al., 2023a; Luan et al., 2024b; Zheng et al., 2024). The perspective taken here is complementary: rather than replacing homophily with another scalar, NSA decomposes node-classification signal into feature, topology, and label-smoothness channels.

Model Interfaces and Pre-Training Diagnosis. Node-classification methods can be grouped by the information interface they expose. *Feature-centric* models such as MLP use node attributes without graph propagation. *Diffusion-based* methods such as LP and APPNP exploit the smoothness

of observed labels over the graph. *Graph-aware* models combine node features with structural context through local aggregation or global attention (Kipf & Welling, 2017; Hamilton et al., 2017; Wu et al., 2023; Rampášek et al., 2022). These interfaces make different assumptions about where discriminative signal lives: features, observed label smoothness, or unlabeled graph structure.

This interface view is the organizing principle behind NSA. Instead of predicting a single winning architecture, NSA measures how much signal is accessible through three broad channels aligned with these interfaces: feature signal, unlabeled topology signal, and label smoothness signal. The result is a pre-training description of the dataset that is more informative than homophily alone and directly useful for model-family triage.

Probing Input-Graph Signal. Linear probes are widely used to quantify what information is accessible from fixed representations (Alain & Bengio, 2016; Tenney et al., 2019). NSA uses probes in a different regime: instead of probing learned neural representations, it probes inexpensive descriptors of the *input graph* itself. These descriptors summarize local feature organization, unlabeled structural roles, and labeled-neighbor smoothness before any graph model is trained. Applying the same fixed probe to each descriptor family yields the three-channel signal profile (FSS, TSS, LSS).

3 Node Signal Atlas

Node classification models differ not just in architecture but in *which information interface* they expose: an MLP reads only features, diffusion methods exploit the smoothness of observed labels along edges, and topology-aware models exploit unlabeled structural context. We introduce the *Node Signal Atlas* (NSA), a diagnostic that measures the strength of these three channels directly from the input graph, *before* training any predictive model.

NSA assigns each dataset three scores. The *Feature Signal Score* (FSS) measures how much label signal is accessible from node features and their local organization in graph-defined neighborhoods; the *Topology Signal Score* (TSS) measures how much is accessible from *unlabeled* structural patterns alone; and the *Label Smoothness Score* (LSS) measures how much is accessible from the smoothness of observed training labels along edges. All three come from the *same fixed* logistic-regression probe evaluated on cheap annulus-based descriptors at hops $k = 0, 1, 2, 3$, under one protocol applied uniformly across datasets.

Annulus neighborhoods. For node v, we use the k-hop *annulus* $A_k(v) = \{u \in V : \mathrm{dist}(u, v) = k\}$ rather than the k-hop ball, separating information by distance and enabling cheap hop-specific summaries at small radii ($k \leq 3$), comparable in reach to a few rounds of message passing.

3.1 Scoring Protocol

For a dataset $D = (G, X, Y)$, each signal family is scored by the same fixed probe – multinomial logistic regression with ℓ_2 regularization, stratified K-fold cross-validation ($K = 5$), and descriptors standardized to zero mean and unit variance *within each training fold*. FSS and TSS descriptors depend only on features and structure and are computed once, reused across folds. LSS descriptors depend on observed labels and must be rebuilt *inside each fold*: a node's label-smoothness descriptor uses only labels available in that fold's inner training partition, so no held-out label ever leaks into its own prediction.

Normalization. We score each probe with macro-averaged, not micro-averaged, accuracy: our benchmarks range from $C = 2$ to $C = 40$ classes, and micro-accuracy's majority-class baseline (which can exceed 0.70 on a skewed binary dataset) makes raw scores incomparable across datasets. Macro-accuracy fixes the chance baseline at $1/C$ regardless of class balance – but $1/C$ itself still swings from 0.50 down to under 0.03 across our suite, so we further normalize: $\widetilde{s} = \dfrac{s - 1/C}{1 - 1/C}$, mapping chance to 0 and perfect classification to 1 for every dataset alike. FSS, TSS, and LSS henceforth refer to this normalized score. Appendix A.5 quantifies the effect of this choice on all 20 real benchmarks.

Computational cost. We cap each annulus $A_k(v)$ at m_k=64 sampled nodes and average over three sampling seeds, keeping cost bounded even on dense graphs. NSA trains no GNN and backpropagates nothing: after one-time descriptor construction, it runs only logistic-regression probes – a cost closer

to training a single standard baseline model than to a full architecture search, and reducible further via node subsampling on very large graphs. Appendix A.6 reports wall-clock time on real benchmarks and the subsampling results for graphs at OGB scale.

3.2 Feature Signal Score (FSS)

FSS measures how much label signal is accessible from node features and their local organization, without using node labels or structural descriptors. It is high when class information is strongly encoded in feature geometry, so feature-centric models are expected to be competitive.

Descriptor. Let $\mathbf{x}_v \in \mathbb{R}^d$ denote the feature vector of node v. We apply a fixed random projection $P \in \mathbb{R}^{r\times d}$ (default $r = 32$), giving $\mathbf{x}_v^{(r)} = P\mathbf{x}_v$. For each annulus $A_k(v) = \{u \in V : \mathrm{dist}(u, v) = k\}$, $k = 0, 1, 2, 3$, we form a hop-specific summary.

For $k = 0$: $\boldsymbol{\mu}_0(v) = \mathbf{x}_v^{(r)}, \quad \boldsymbol{\sigma}_0(v) = \mathbf{0}, \quad \mathrm{shift}_0(v) = 0.$

For $k \geq 1$, let $n_k(v) = |A_k(v)|$. If $n_k(v) = 0$, we zero-pad the entire kth hop block. Otherwise:

$$\boldsymbol{\mu}_k(v) = \frac{1}{n_k(v)} \sum_{u\in A_k(v)} \mathbf{x}_u^{(r)}, \quad \boldsymbol{\sigma}_k(v) = \sqrt{\frac{1}{n_k(v)} \sum_{u\in A_k(v)} \left(\mathbf{x}_u^{(r)} - \boldsymbol{\mu}_k(v)\right)^{\odot 2}}, \quad \mathrm{shift}_k(v) = \|\boldsymbol{\mu}_k(v) - \mathbf{x}_v^{(r)}\|_2.$$

The center-shift captures how atypical the node is relative to the feature cloud at hop k; large values indicate feature heterophily at that radius. Concatenating four hop blocks gives:

$$\boldsymbol{\Phi}_F(v) = \left[\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \mathrm{shift}_0 \;\middle|\; \cdots \;\middle|\; \boldsymbol{\mu}_3, \boldsymbol{\sigma}_3, \mathrm{shift}_3\right] \in \mathbb{R}^{4(2r+1)}.$$

3.3 Topology Signal Score (TSS)

TSS measures how much label signal is accessible from *unlabeled* graph structure, using neither node features nor class labels. It is high when local ego-structures exhibit informative role patterns, core–periphery contrast, community-like closure, or distinctive growth behavior, that correlate with labels. In such cases, graph-aware models exploiting structural information are most likely to provide gains.

Descriptor. We precompute three base per-node statistics: **(i)** log-degree $\log(\deg(v)+1)$, **(ii)** core number $\mathrm{core}(v)$, and **(iii)** egonet density $\mathrm{ego_dens}(v) = \dfrac{2\,E(N(v))}{\deg(v)(\deg(v)-1)+\varepsilon}$, where $E(N(v))$ counts edges among v's neighbors and $\varepsilon = 10^{-8}$.

The $k = 0$ block is simply these three statistics evaluated at v itself:

$$\boldsymbol{\phi}_S(v, 0) = \left[\log(\deg(v) + 1),\ \mathrm{core}(v),\ \mathrm{ego_dens}(v)\right] \in \mathbb{R}^3.$$

For $k = 1, 2, 3$, let $n_k(v) = |A_k(v)|$ and $b_k(v) = \sum_{i=0}^{k} n_i(v)$; if $n_k(v) = 0$ we zero-pad the entire block. Otherwise, we aggregate over the annulus $A_k(v)$ along four axes:

- *Size:* $\log(n_k+1)$ and $\log(b_k+1)$, the annulus and cumulative ball sizes.
- *Degree:* mean and std. of $\log(\deg(u)+1)$ over $u \in A_k(v)$, $\bar{d}_k = \frac{1}{n_k}\sum_u \log(\deg(u)+1)$ and $\sigma_{d,k} = \sqrt{\frac{1}{n_k}\sum_u (\log(\deg(u)+1) - \bar{d}_k)^2}$.
- *Expansion:* the frontier expansion ratio $\rho_k = n_k/\max(n_{k-1}, 1)$ and surface-to-volume ratio $s_k = n_k/(b_k+1)$, which together distinguish hub-dominated neighborhoods from uniform expansion.
- *Role:* annulus means of core number and egonet density, $\overline{\mathrm{core}}_k = \frac{1}{n_k}\sum_u \mathrm{core}(u)$ and $\overline{\mathrm{ego}}_k = \frac{1}{n_k}\sum_u \mathrm{ego_dens}(u)$, together with λ_k, the Weisfeiler–Lehman role-diversity term defined next.

Concatenating these gives the 9-dimensional block

$$\boldsymbol{\phi}_S(v, k) = \left[\log(n_k+1),\ \bar{d}_k,\ \sigma_{d,k},\ \rho_k,\ s_k,\ \overline{\mathrm{core}}_k,\ \overline{\mathrm{ego}}_k,\ \lambda_k,\ \log(b_k+1)\right] \in \mathbb{R}^9.$$

WL role diversity. To capture structural roles beyond degree and growth, we run T rounds of global 1-WL refinement initialized by binned log-degree $\lfloor \log_2(\deg+1) \rfloor$, hash colors into B buckets (default $B = 1024$), and compute normalized entropy over each annulus.

For $n_k(v) = 0$ we set $\lambda_k(v) = 0$. Otherwise: $\lambda_k(v) = \frac{1}{T}\sum_{t=1}^{T} \frac{H(p_{v,k,t})}{\log(n_k(v)+1)}$, $p_{v,k,t}(a) = \frac{|\{u \in A_k(v) : h(\mathrm{wl}_t(u)) = a\}|}{n_k(v)}$, where $h(\cdot)$ maps colors to $\{1, \ldots, B\}$. Higher entropy indicates richer structural heterogeneity within the annulus; lower entropy indicates a more structurally uniform neighborhood.

The full topology descriptor concatenates all hop blocks:

$$\mathbf{\Phi}_S(v) = \left[\boldsymbol{\phi}_S(v,0) \mid \boldsymbol{\phi}_S(v,1) \mid \boldsymbol{\phi}_S(v,2) \mid \boldsymbol{\phi}_S(v,3)\right] \in \mathbb{R}^{30}.$$

Descriptor standardization is especially important here because components span very different scales: $\mathrm{ego_dens} \in [0,1]$ while log-degree is unbounded, and without standardization degree features dominate the probe.

3.4 Label Smoothness Score (LSS)

LSS measures how much label signal is accessible from the smoothness of observed training labels across the graph. It is high when labeled neighbors are class-consistent, so diffusion-style methods such as LP and APPNP can exploit the graph effectively. Because LSS uses only labels available within each training fold, it is computed without label leakage.

Descriptor. For a given cross-validation fold, let $V_{\mathrm{train}}^{(\mathrm{fold})}$ denote the labeled nodes in that fold's inner training partition. For each node v and $k = 1, 2, 3$, define the labeled annulus $A_k^{\mathrm{lab}}(v) = A_k(v) \cap V_{\mathrm{train}}^{(\mathrm{fold})}$, with $n_k(v) = |A_k(v)|$ and $m_k(v) = |A_k^{\mathrm{lab}}(v)|$. The empirical class distribution over labeled annulus members is $\mathbf{p}_k(v)[c] = \frac{|\{u \in A_k^{\mathrm{lab}}(v) : y_u = c\}|}{\max(m_k(v), 1)}$, set to uniform when $m_k(v) = 0$. Two coverage features accompany each distribution: $\mathrm{cov}_k(v) = \frac{m_k(v)}{n_k(v)+1}$, $\ell_k(v) = \log(m_k(v) + 1)$.

The coverage fraction cov_k disambiguates a nearly uniform label histogram arising from genuine label mixing from one arising from an absence of labeled neighbors, a critical distinction in semi-supervised settings. We omit $k = 0$ since $A_0(v) = \{v\}$ would trivially expose the node's own label. Concatenating three hop blocks:

$$\mathbf{\Phi}_H(v) = \left[\mathbf{p}_1, \mathrm{cov}_1, \ell_1 \mid \mathbf{p}_2, \mathrm{cov}_2, \ell_2 \mid \mathbf{p}_3, \mathrm{cov}_3, \ell_3\right] \in \mathbb{R}^{3(C+2)}.$$

Atlas embedding and interpretation. Applying the scoring protocol and normalization yields the atlas embedding (FSS, TSS, LSS), where 0 denotes chance and 1 perfect recovery. Scores below 0 indicate descriptors that mislead the probe; in practice such values are rare and close to 0. We treat a normalized score above $\delta = 0.05$ as practically active. The four canonical regions are: **high FSS** (feature-dominant, MLP competitive), **high LSS** (smoothness-dominant, LP and APPNP competitive), **high TSS** (topology-dominant, graph-aware models benefit most), and **all scores near** 0 (signal desert, all families struggle). When multiple scores are elevated, atlas position captures their co-activation, giving a richer characterization than homophily, which reflects at most the smoothness dimension.

4 Experiments

We evaluate NSA through three complementary analyses. First, on controlled synthetic data, we verify that FSS, TSS, and LSS respond selectively to their intended channels and correctly expose the limits of marginal measurement (Section 4.2; full construction and results in Appendix B). Second, on 20 real benchmarks, we show that atlas position reveals signal regimes that are not visible from homophily alone (Section 4.3; extended analysis in Appendix C). Third, on a separate frozen cross-generator benchmark purpose-built to test architecture selection, we show that NSA's coordinates carry selection-relevant information beyond simple homophily statistics and combine productively with richer homophily-based ones, rather than replacing them (Section 4.4; full construction, protocol, and robustness checks in Appendix D). Together, these analyses support one thesis: *signal composition, not homophily alone, is the right lens for understanding when and why different node-classification model interfaces succeed.* Appendix C.4 gives a detailed comparison with related graph characterization frameworks.

4.1 SETUP

Datasets. We evaluate on 20 real-world node classification benchmarks and 24 controlled synthetic datasets. The real benchmarks span heterophilic, mixed, and homophilic graphs drawn from webpage, social, citation, and co-purchase networks Pei et al. (2019); Rozemberczki et al. (2021); Platonov et al. (2023b); McCallum et al. (2000); Shchur et al. (2018). Dataset statistics and NSA scores are in Appendix A.1; the synthetic suite is described in Appendix B. A third, separate 270-problem cross-generator benchmark is used only for the architecture-triage study in Section 4.4 and is described in full in Appendix D.

Baselines. We evaluate nine models spanning four model families; see Appendix A.2 for full descriptions.

- **F** - Feature-only: MLP
- **L** - Label propagation/diffusion: LP Zhu et al. (2003), APPNP Gasteiger et al. (2019)
- **T** - Topology-aware GNNs: GCN Kipf & Welling (2017), GSAGE Hamilton et al. (2017), H2GCN Zhu et al. (2020), SGC Wu et al. (2019)
- **GT** - Graph transformers: GPS Rampášek et al. (2022), SGFormer Wu et al. (2023)

The architecture-triage study in Section 4.4 uses a separate, smaller bank of six architectures with more distinct signal-access mechanisms (MLP, GCN, GraphSAGE, APPNP, H2GCN, GIN); we describe that choice and its motivation in Appendix D.

Evaluation and implementation. Model performance is micro-accuracy for all multi-class datasets. NSA scores use normalized macro-accuracy throughout: $\widetilde{s} = (s - 1/C)/(1 - 1/C)$, so 0 = chance and 1 = perfect, enabling cross-dataset comparison regardless of C; this choice is justified in detail in Appendix A.5. All models are evaluated using stratified 60/20/20 train/validation/test splits, except on ogbn-arxiv, where we use OGB's official split. We report mean accuracy over 10 runs. NSA scores are averaged over three sampling seeds. All GNNs use two layers of width 64, trained up to 500 epochs with early stopping (patience 50). Hyperparameters are selected by grid search over learning rate, weight decay, and dropout (12 combinations); see Appendix A.4 for the full grid and Appendix A.3 for the evaluation and NSA scoring protocol. NSA parameters are fixed across all datasets: $r = 32$, $T = 4$, $m_k = 64$. NSA is computed once per dataset before any downstream model training; wall-clock cost and a large-graph subsampling strategy are reported in Appendix A.6. Our implementation is available at
https://anonymous.4open.science/r/NodeAtlas-C317

4.2 SYNTHETIC VALIDATION: CALIBRATION AND SCOPE

Because NSA is a diagnostic instrument rather than a predictive model, the synthetic study is used to validate measurement under controlled signal manipulations, not to establish state-of-the-art performance. We therefore construct a 24-dataset synthetic suite in which feature signal, unlabeled topology signal, and label smoothness can be varied systematically: the 8 corners of the $2 \times 2 \times 2$ F/T/L cube, plus 16 intermediate datasets that populate the interior of the atlas. Construction details and full results are given in Appendix B.

The synthetic study is designed to test three properties: *selectivity* (does the corresponding NSA score rise when one channel is activated and the others are suppressed?), *calibration/compositionality* (do mixed-signal datasets move to the expected interior regions of the atlas?), and *scope* (what happens when labels are recoverable only through interactions outside NSA's three marginal probes?).

Table 1 reports the 8 corner datasets. These already support all three properties. In the pure corners F, T, and L, the score associated with the active channel rises well above chance. The main exception is the T corner (TSS $= 0.711$, LSS $= 0.786$), where topology and smoothness co-occur by construction; this coupling is asymmetric (elevated topology signal induces elevated smoothness, but not the reverse) and is discussed in full in Appendix B.1. Even in the T corner, graph-aware models are the clear winners, confirming that the atlas still identifies the relevant interface despite the coupling.

The mixed corners FT, FL, TL, and FTL confirm compositionality: multiple scores rise together in the expected pattern, and model behavior reflects the mixture. Finally, the Node Twins dataset C serves as a controlled boundary case. All three marginal NSA scores are near chance, yet GPS reaches 87.95%, 34.22 points above the next-best model, and SGFormer reaches 53.73% – two architecturally distinct

Table 1: **Synthetic corner datasets.** NSA scores and model performance on the 8 corner cases of the $2 \times 2 \times 2$ F/T/L cube ($N = 1500$, $C = 3$); chance = 33.3%). FSS, TSS, and LSS are normalized macro-accuracy NSA scores (0 = chance, 1 = perfect). **Best**, **second**, and **third** performances in each row are highlighted. Full results for 24 synthetic datasets are reported in Appendix B.

Dataset	FSS	TSS	LSS	MLP	LP	APPNP	GCN	SAGE	H2GCN	SGC	GPS	SGFormer
F	0.681	−0.012	0.001	**86.70**	33.99	78.48	35.68	56.83	62.61	64.22	**85.54**	**85.87**
T	0.257	0.711	0.786	31.77	33.37	33.70	**68.10**	52.53	65.90	36.50	**74.97**	**74.93**
L	0.251	0.142	0.773	31.77	**89.43**	50.70	62.03	55.17	65.93	49.23	**71.93**	**79.37**
FT	0.864	0.711	0.786	**85.57**	33.37	76.50	70.40	74.67	82.27	83.40	**93.83**	**94.60**
FL	0.899	0.142	0.773	85.57	89.43	**95.83**	72.03	78.03	83.83	94.67	**94.97**	**95.37**
TL	0.520	0.928	0.911	31.77	79.60	55.13	70.97	62.23	**82.43**	48.17	**86.37**	**85.47**
FTL	0.927	0.928	0.911	85.57	79.60	**95.93**	81.53	78.33	90.00	94.60	**96.13**	**97.17**
C	−0.016	−0.014	-0.018	32.67	33.14	**38.61**	33.80	33.14	32.74	38.15	**87.95**	**53.73**

models relying on global or joint graph-feature processing both clearing chance. This occurs because labels are encoded through a joint interaction between feature type and structural role, which lies outside NSA's marginal scope but can be partially recovered by positional and joint processing. Thus, C is not a failure case of NSA, but an explicit illustration of where marginal measurement stops; we return to this scope boundary in Appendix B.2.

Together, these synthetic results justify using NSA as a calibrated diagnostic lens on real node-classification benchmarks, while the intermediate datasets in Appendix B.3 show that atlas positions vary smoothly between corner regimes rather than collapsing onto a few discrete labels.

4.3 Real Benchmark Atlas: Signal Regimes Beyond Homophily

Having validated NSA under controlled synthetic signal manipulations, we now apply it to real node classification benchmarks. The goal is to test whether the calibrated atlas reveals meaningful structure in real data. Table 3 reports NSA scores together with model performance on all 20 benchmarks, and Figure 1 (in Appendix) plots the same datasets in atlas space; full results with standard deviations and the corresponding dataset statistics are given in Appendix C.1 and Appendix A.1. Once projected into (FSS, TSS, LSS) space, the benchmark suite is no longer a flat list of names and accuracies, but a structured landscape of accessible signal regimes.

Atlas geometry. Several broad patterns emerge immediately. High-homophily citation and co-purchase datasets such as Cora, PubMed, Amazon-Computers, and Amazon-Photo occupy the high-LSS region, consistent with reliable label propagation and strong diffusion or graph-aware performance. At the same time, low homophily does not imply weak accessible signal: Roman-Empire and Texas have very small h_{node}, yet both exhibit strong FSS and strong model performance, showing that *heterophilic* does not mean *signal-poor*.

At the opposite extreme, Actor lies close to a signal desert, with all three scores near zero (FSS = 0.120, TSS = 0.004, LSS = 0.007) and tightly clustered accuracies, indicating that no single marginal channel is strongly informative. Conversely, Genius shows why chance-normalization matters at the atlas level: despite moderate homophily ($h_{\text{node}} = 0.509$) and high raw accuracies (82–86% across families), all three normalized coordinates are near zero ($0.037, 0.070, 0.086$), correctly indicating that its apparent easiness reflects an 80% majority class rather than accessible class-discriminative signal (Appendix A.5). Thus, real benchmarks are not well described by a single homophily axis, but occupy a genuinely three-dimensional signal space. Appendix C.1 gives a fuller reading of these benchmark patterns and discusses the complementary role of TSS in detail (Appendix C.3).

Table 2: Dataset pairs with similar h_{node}; the last column shows best accuracy. See Table 13 for details.

Dataset	h_{node}	FSS	TSS	LSS	Acc.
Flickr	0.243	0.241	0.162	0.343	92.70
Chameleon†	0.244	0.199	0.234	0.140	41.15
Cornell	0.200	0.298	0.097	0.105	75.75
Actor	0.202	0.120	0.004	0.007	37.57
Cornell	0.200	0.298	0.097	0.105	75.75
Squirrel†	0.191	0.102	0.201	0.071	39.69
Chameleon	0.247	0.511	0.321	0.282	63.18
Chameleon†	0.244	0.199	0.234	0.140	41.15

Beyond homophily. The main value of the atlas is that it separates datasets that appear similar under homophily but behave differently in practice. Homophily is a useful first-order descriptor,

Table 3: **Real Benchmarks.** Node classification mean accuracy on 20 real benchmarks, ordered by node homophily h_{node}. FSS, TSS, LSS are NSA scores with 0 = chance, 1 = perfect. **Best**, **second** and **third** results are highlighted. Full results with standard deviations are given in Appendix C.1.

Dataset	h_{node}	FSS	TSS	LSS	MLP	LP	APPNP	GCN	SAGE	H2GCN	SGC	GPS	SGFormer
Roman-Empire	0.046	0.537	0.131	0.203	66.12	8.08	56.33	36.56	66.30	**70.84**	41.75	**82.75**	**80.07**
Texas	0.056	0.415	0.411	0.474	78.78	19.27	51.71	55.37	74.15	**80.00**	50.24	**80.73**	**80.49**
Wisconsin	0.099	0.337	0.139	0.150	**86.42**	17.92	51.32	49.43	**82.08**	81.70	46.04	**82.26**	**86.42**
Squirrel†	0.191	0.102	0.201	0.071	35.41	32.94	32.74	**37.93**	33.25	32.96	**38.02**	**39.69**	36.24
Cornell	0.200	0.298	0.097	0.105	**75.75**	23.25	47.50	42.25	71.25	69.25	42.25	**75.25**	**75.50**
Actor	0.202	0.120	0.004	0.007	**36.73**	18.36	29.00	29.19	**37.15**	36.70	28.73	**37.57**	36.51
Squirrel	0.215	0.286	0.200	0.125	34.57	16.57	29.32	28.62	33.11	**43.50**	27.39	**35.70**	**35.51**
Flickr	0.243	0.241	0.162	0.343	91.40	41.34	92.17	70.81	**92.33**	**92.70**	67.84	90.52	**92.32**
Chameleon†	0.244	0.199	0.234	0.140	**39.18**	33.28	38.42	36.17	35.85	36.23	36.23	**41.15**	**39.34**
Chameleon	0.247	0.511	0.321	0.282	51.33	21.15	43.64	42.88	49.83	**63.18**	38.58	**52.22**	**52.88**
Amazon-Ratings	0.376	0.076	0.024	0.149	44.27	**46.98**	46.13	43.37	43.20	45.49	41.94	**48.81**	**50.88**
BlogCatalog	0.391	0.466	0.288	0.594	93.18	65.34	95.57	77.93	95.56	**96.38**	78.19	**95.89**	**96.30**
Genius	0.509	0.037	0.070	0.086	82.55	80.36	81.38	82.43	**84.49**	**84.86**	81.91	OOM	**85.55**
Deezer-Europe	0.530	0.091	0.010	0.060	**67.47**	56.51	65.60	61.99	66.76	**67.33**	62.97	66.89	**67.51**
Ogbn-arxiv	0.635	0.413	0.059	0.481	52.48	**67.62**	**68.48**	59.86	60.75	63.76	65.84	OOM	**70.47**
CiteSeer	0.706	0.399	0.182	0.552	72.83	70.64	**77.21**	76.86	**77.45**	**77.18**	76.81	76.35	76.32
PubMed	0.792	0.678	0.315	0.711	87.83	83.20	88.27	87.68	88.93	**89.63**	84.78	**89.34**	**89.91**
Am.-Computers	0.800	0.728	0.342	0.835	85.34	89.35	**90.70**	88.57	88.13	90.31	89.67	**90.91**	**91.54**
Cora	0.825	0.647	0.287	0.766	74.74	84.87	**88.81**	87.01	87.50	**87.94**	**87.72**	84.60	85.42
Amazon-Photo	0.850	0.823	0.577	0.876	92.40	91.64	**95.19**	93.54	**95.21**	95.09	93.76	94.68	**95.69**

but it is most closely aligned with the smoothness channel and does not directly capture feature accessibility or unlabeled structural-role signal. As a result, datasets with nearly identical h_{node} can occupy very different positions in (FSS, TSS, LSS) space and favor different model families. For example, Flickr ($h_{\text{node}} = 0.243$) and Chameleon† ($h_{\text{node}} = 0.244$) differ by only $\Delta h = 0.001$, yet are separated by 51.6 percentage points in best-model accuracy; NSA correctly identifies the source as a difference in smoothness signal (LSS $= 0.343$ vs. 0.140) that homophily cannot detect. Similarly, Cornell ($h_{\text{node}} = 0.200$, FSS $= 0.298$) and Actor ($h_{\text{node}} = 0.202$, FSS $= 0.120$) are nearly indistinguishable by homophily yet differ by 38 percentage points in best-model accuracy. Thus, the atlas does not replace the homophily view; it empirically refines it into a more informative signal-based characterization. Two additional matched-homophily pairs, including a case where the two datasets are distinguished not by signal *strength* but by which channel dominates, are given in Appendix C.2; Appendix C.4 further relates this reading of the raw/filtered Chameleon pair to recent evaluation-rigor critiques of homophily-based benchmarking (Luan et al., 2024c).

Representative benchmark regimes. The atlas also explains several recurring benchmark phenomena. Feature-accessible datasets explain why MLP can remain competitive despite weak homophily: on the six datasets with $h_{\text{node}} < 0.30$ and FSS > 0.25, MLP outperforms GCN by an average of 23.0 percentage points. High FSS should nonetheless be read as *include MLP in the shortlist* rather than *MLP must win*, since datasets such as Texas and Chameleon show that strong feature signal can coexist with competitive graph-aware models. Signal-desert datasets explain why some benchmarks remain difficult across all model families: the issue is not simply low homophily, but the absence of a strong marginal signal through any one of the three interfaces measured by NSA.

Overall, the real-benchmark atlas shows that the calibrated synthetic measurement transfers meaningfully to real data. Real benchmarks occupy distinct and interpretable signal regimes, and those regimes explain recurring model behavior more richly than homophily alone. This motivates the next subsection, where we quantify the added value of NSA beyond homophily on a separate, frozen architecture-selection benchmark (Section 4.4; full construction and protocol in Appendix D).

4.4 Quantitative Utility of NSA

Architecture Triage: NSA Complements Homophily. Although NSA is primarily a characterization framework rather than a model selector, its coordinates should be useful for coarse architecture triage if they capture distinct sources of predictive signal. We test this on a frozen cross-generator benchmark containing six synthetic signal regimes and six specialized architectures (MLP, GCN, GraphSAGE, APPNP, H2GCN, and GIN), chosen specifically because they expose more distinct signal-access mechanisms than the flexible generalists (GPS, SGFormer) used elsewhere in this paper. The benchmark was calibrated using only architecture-performance diversity and selection headroom, and the standalone representations below were evaluated only after this freeze; full construction details, the pilot–freeze–final protocol, and robustness checks are given in Appendix D.

Table 4 shows that NSA substantially reduces regret relative to node homophily (from 6.89 to 2.63 pp). Both representations also outperform *constant selector*, a baseline that always recommends the training corpus's best-performing architecture regardless of the input graph (7.18 pp regret), confirming genuine graph-specific signal rather than a fixed default preference. The richer Homophily-4 profile achieves lower regret and higher Top-1 accuracy than NSA alone, while NSA is competitive on Spearman correlation and stronger on Top-3 coverage; the two representations are therefore not interchangeable.

Table 4: **Architecture selection performance.** Regret (in percentage points, pp) is the primary metric. All learned selectors use the same standardized multi-output Ridge predictor.

Representation	Dim.	Regret (pp) ↓	Spearman ↑	Top-1 ↑	Top-3 ↑
Constant Selector	–	7.18	0.484	0.422	0.867
Node Homophily	1	6.89	0.468	0.422	0.822
Homophily-4	4	1.62	0.679	0.600	0.811
NSA	3	2.63	0.682	0.556	0.889
NSA + Node Hom.	4	**1.18**	0.766	**0.678**	0.867
NSA + Homophily-4	7	1.31	**0.788**	0.656	**0.900**

Combining them improves triage further, and the two combinations serve different practical purposes. NSA + Node Homophily attains the lowest regret and the highest Top-1 accuracy – the better choice when a single evaluation must be committed to. NSA + Homophily-4 instead gives the strongest overall ranking, with the highest Spearman correlation and Top-3 coverage – the better choice when a short list is acceptable, consistent with the budgeted-search results in Appendix D.6. In particular, augmenting Homophily-4 with NSA reduces regret from 1.62 to 1.31 pp ($p = 0.0234$, paired Wilcoxon). Thus, NSA need not replace conventional label–graph statistics; its feature-, topology-, and label-context coordinates provide complementary information about *where* useful signal is accessible. This complements the real-benchmark stress test in Appendix D.7: dataset diagnostics retain meaningful, non-trivial selection headroom precisely when candidate architectures expose sufficiently distinct signal-access mechanisms, which is also why that headroom compresses on the 20 real benchmarks once flexible generalist architectures are included.

Descriptive Correlations on the Real Benchmarks. Returning to the 20 real benchmarks, we examine Pearson and Spearman correlations between NSA scores and model performance as a complementary sanity check: if the three coordinates are genuinely aligned with the model interfaces they are meant to characterize, that alignment should be visible directly in per-model accuracy, not only in the triage result above. LSS is the strongest and most consistent predictor across all model families ($r \in [0.595, 0.771]$, all $p < 0.01$), while FSS significantly predicts both feature-centric and graph-aware performance. TSS has weaker marginal correlations with raw accuracy than FSS and LSS, consistent with its role as a conditional structural axis rather than a standalone performance predictor: Appendix C.3 shows that elevated TSS is nonetheless informative for the more targeted question of when topology-aware models gain over MLP, and reports the full interface- and model-level correlation tables.

Table 5: Pearson correlations between normalized NSA scores and representative model test accuracy across 20 real benchmarks.

	MLP	APPNP	H2GCN	GPS
FSS	0.510^{*}	0.557^{*}	0.616^{**}	0.690^{**}
TSS	0.408	0.390	0.458^{*}	0.453
LSS	0.595^{**}	0.771^{**}	0.700^{**}	0.730^{**}

5 Conclusion

We introduced the *Node Signal Atlas* (NSA), a lightweight pre-training framework that characterizes node-classification problems through three complementary signal channels: node features, unlabeled structural roles, and local label context. Its central premise is that graph datasets should be characterized not only by how labels align with edges, but by *where predictive information is accessible*. Across controlled and real benchmarks, NSA reveals distinctions that homophily alone obscures. On a separate architecture-selection benchmark, NSA's coordinates further prove complementary to homophily-based statistics, improving low-budget model triage when combined. Throughout, the two views are complementary rather than competing: homophily captures label–adjacency alignment, while NSA describes accessible feature, structural, and local label-context signal. NSA is not an architecture oracle and does not detect signal that depends on interactions between channels, but it provides an interpretable first-pass description of the learning problem before model selection. More broadly, this signal-centered view offers a richer alternative to single-scalar graph characterization.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and revising portions of the manuscript, primarily to improve English, clarity, organization, and readability; to generate and edit research code; to implement author-specified procedures for generating the synthetic datasets used in the experiments; and to help create and refine the pipeline figure. The synthetic mechanisms, parameters, and experimental design were specified by the authors, while AI-assisted code was reviewed, tested, and verified before use, and the resulting datasets and experimental outputs were checked for consistency with the intended constructions. We did not use generative AI to independently develop the conceptual framework, formulate the NSA descriptors, or interpret experimental results. All AI-assisted text, code, data-generation procedures, and visual material were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies methodological questions in graph learning and does not involve human subjects, private or personally identifiable information, or deployment in high-risk settings. The real-world benchmark datasets used in our experiments are publicly available research datasets, and our synthetic datasets are generated solely for controlled methodological evaluation. We do not release sensitive data or make claims about protected attributes, individuals, or populations. As with any dataset-characterization or model-selection methodology, NSA could in principle be applied in downstream domains with domain-specific fairness, privacy, or safety concerns; such use would require separate validation and appropriate safeguards. We are not aware of additional ethical concerns specific to the methods or experiments presented here.

REPRODUCIBILITY STATEMENT

We provide complete implementation code and experiment scripts in an anonymous repository, including NSA descriptor construction and scoring, synthetic-data generation, downstream architecture training, the architecture-selection experiments, ablations, and figure/table generation. The main paper describes the methodology and evaluation protocol (Sections 3 and 4), while Appendix A gives the dataset suite, baseline configurations, evaluation protocol, and NSA scoring details (including the fixed probe, normalization, and hyperparameters r, T, m_k); Appendix B specifies the synthetic generators and signal controls; and Appendix D documents the cross-generator architecture-selection benchmark, its pilot–freeze–final protocol, and the associated robustness checks. Computational cost and the large-graph subsampling procedure are reported in Appendix A.6. All reported experiments use fixed random seeds and the stratified train/validation/test protocol described in the appendix. Together, the released code and appendix are intended to reproduce the synthetic benchmarks, real-data experiments, ablations, and all reported results. Our implementation is available at
https://anonymous.4open.science/r/NodeAtlas-C317

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *NeurIPS*, 30, 2017.

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

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *NeurIPS*, 34:20887–20902, 2021.

Sitao Luan, Chenqing Hua, Qincheng Lu, Liheng Ma, Lirong Wu, Xinyu Wang, Minkai Xu, Xiao-Wen Chang, Doina Precup, Rex Ying, et al. The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges. *arXiv preprint arXiv:2407.09618*, 2024a.

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. *NeurIPS*, 36, 2024b.

Sitao Luan, Qincheng Lu, Chenqing Hua, Xinyu Wang, Jiaqi Zhu, Xiao-Wen Chang, Guy Wolf, and Jian Tang. Are heterophilic GNNs and homophily metrics really effective? Evaluation pitfalls and new benchmarks. *arXiv preprint arXiv:2409.05755*, 2024c.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? In *ICLR*, 2022.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *ICLR*, 2019.

Oleg Platonov, Denis Kuznedelev, Artem Babenko, and Liudmila Prokhorenkova. Characterizing graph datasets for node classification: Homophily-heterophily dichotomy and beyond. *Advances in Neural Information Processing Systems*, 36:523–548, 2023a.

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? *ICLR*, 2023b.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1325–1334, 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. *Relational Representation Learning Workshop, NeurIPS 2018*, 2018.

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 807–816, 2009.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *ICML*, pp. 6861–6871. PMLR, 2019.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. Sgformer: Simplifying and empowering transformers for large-graph representations. *Advances in Neural Information Processing Systems*, 36:64753–64773, 2023.

Yilun Zheng, Sitao Luan, and Lihui Chen. What is missing for graph homophily? disentangling graph homophily for graph neural networks. *Advances in Neural Information Processing Systems*, 37:68406–68452, 2024.

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 *NeurIPS*, 2020.

Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 912–919. AAAI Press, 2003.

APPENDIX

A Experimental Details **13**

A.1 Datasets . . . 13

A.2 Baseline Models . . . 14

A.3 Evaluation and NSA Scoring Protocol . . . 14

A.4 Implementation Details . . . 15

A.5 Why NSA Uses Normalized Macro-Accuracy . . . 15

A.6 Computational Cost . . . 16

B Extended Analysis on Synthetic Benchmarks **17**

B.1 Construction and Signal Controls . . . 17

B.2 Corner and Boundary Cases . . . 18

B.3 Intermediate Sweep and Full Results . . . 18

B.4 Guidance Rule . . . 19

C Extended Analysis on Real Benchmarks **20**

C.1 Full Results and the Real-Data Atlas . . . 20

C.2 Beyond Homophily: Similar Homophily, Different Problems . . . 21

C.3 TSS as a Conditional Signal . . . 22

C.4 Comparison with Related Characterization Approaches . . . 23

D Architecture Triage **25**

D.1 Cross-Generator Architecture-Selection Benchmark . . . 25

D.2 Pilot–Freeze–Final Protocol . . . 26

D.3 Benchmark Validity . . . 27

D.4 NSA and Homophily Are Complementary . . . 27

D.5 Where Does the Complementarity Come From? . . . 28

D.6 Budgeted Architecture Search . . . 29

D.7 Generalist Architectures Reduce Selection Headroom . . . 29

A EXPERIMENTAL DETAILS

A.1 DATASETS

We evaluate NSA on 20 node-classification benchmarks spanning a broad range of graph sizes, domains, and homophily levels. The suite includes the WebKB graphs Cornell, Texas, and Wisconsin (Pei et al., 2019); Chameleon and Squirrel (Rozemberczki et al., 2021); Roman-Empire, Actor, Flickr, Chameleon-filtered, Squirrel-filtered, Amazon-Ratings, and BlogCatalog (Platonov et al., 2023b); the citation networks Cora, CiteSeer, and PubMed (McCallum et al., 2000); Amazon-Computers and Amazon-Photo (Shchur et al., 2018); Deezer-Europe (Rozemberczki & Sarkar, 2020); and two large-scale datasets, Genius (Lim et al., 2021) and Ogbn-arxiv (Hu et al., 2020).

Table 6 reports the dataset statistics and NSA coordinates used throughout the paper. Datasets are ordered by node homophily to make clear that low-homophily graphs occupy several distinct regions of NSA space rather than forming a single regime.

Table 6: **Dataset statistics and NSA scores.** C = # classes, d = #features, h_{node} = node homophily; h_{edge} = edge homophily. Datasets ordered by h_{node}.

Dataset	# nodes	# edges	C	d	h_{node}	h_{edge}	FSS	TSS	LSS
Roman-Empire (Platonov et al., 2023b)	22,662	32,927	18	300	0.046	0.047	0.537	0.131	0.203
Texas (Pei et al., 2019)	183	309	5	1,703	0.056	0.062	0.415	0.411	0.474
Wisconsin (Pei et al., 2019)	251	499	5	1,703	0.099	0.170	0.337	0.139	0.150
Squirrel-filtered (Platonov et al., 2023b)	2,223	46,998	5	2,089	0.191	0.207	0.102	0.201	0.071
Cornell (Pei et al., 2019)	183	295	5	1,703	0.200	0.298	0.298	0.097	0.105
Actor (Tang et al., 2009)	7,600	33,544	5	931	0.202	0.216	0.120	0.004	0.007
Squirrel (Rozemberczki et al., 2021)	5,201	217,073	5	2,089	0.215	0.223	0.286	0.200	0.125
Flickr (Platonov et al., 2023b)	7,575	239,738	9	12,047	0.243	0.239	0.241	0.162	0.343
Chameleon-filtered (Platonov et al., 2023b)	890	8,854	5	2,325	0.244	0.236	0.199	0.234	0.140
Chameleon (Rozemberczki et al., 2021)	2,277	36,101	5	2,325	0.247	0.234	0.511	0.321	0.282
Amazon-Ratings (Platonov et al., 2023b)	24,492	93,050	5	300	0.376	0.380	0.076	0.024	0.149
BlogCatalog (Platonov et al., 2023b)	5,196	171,743	6	8,189	0.391	0.401	0.466	0.288	0.594
Genius (Lim et al., 2021)	421,961	1,845,736	2	12	0.509	0.593	0.037	0.070	0.086
Deezer-Europe (Rozemberczki & Sarkar, 2020)	28,281	92,752	2	31,241	0.530	0.525	0.091	0.010	0.060
Ogbn-arxiv (Hu et al., 2020)	169,343	2,315,598	40	128	0.635	0.654	0.413	0.059	0.481
CiteSeer (McCallum et al., 2000)	3,327	4,732	6	3,703	0.706	0.736	0.399	0.182	0.552
PubMed (McCallum et al., 2000)	19,717	44,338	3	500	0.792	0.802	0.678	0.315	0.711
Amazon-Computers (Shchur et al., 2018)	13,752	245,861	10	767	0.800	0.780	0.728	0.342	0.835
Cora (McCallum et al., 2000)	2,708	5,429	7	1,433	0.825	0.810	0.647	0.287	0.766
Amazon-Photo (Shchur et al., 2018)	7,650	119,081	8	745	0.850	0.830	0.823	0.577	0.876

A.2 Baseline Models

We evaluate nine downstream models covering the principal interfaces relevant to the NSA coordinates.

Feature-centric. **MLP** (Goodfellow et al., 2016) operates independently on node features and ignores graph structure.

Diffusion and propagation. **LP** (Zhu et al., 2003) directly propagates observed labels over the graph. **APPNP** (Gasteiger et al., 2019) instead propagates predictions produced from node features through personalized PageRank.

Message-passing GNNs. **GCN** (Kipf & Welling, 2017) uses symmetrically normalized local aggregation; **GraphSAGE** (Hamilton et al., 2017) uses mean neighborhood aggregation with a separate ego contribution; **H2GCN** (Zhu et al., 2020) separates ego and neighborhood representations and incorporates higher-order neighborhoods; and **SGC** (Wu et al., 2019) collapses repeated graph propagation into a linear prediction model.

Graph transformers. **GPS** (Rampášek et al., 2022) combines local message passing with global self-attention and positional encodings, while **SGFormer** (Wu et al., 2023) uses a simplified global-attention architecture.

These models are not intended to form a strict one-to-one correspondence with the three NSA axes. Rather, they provide a diverse collection of interfaces whose behavior can be interpreted in light of feature accessibility, structural information, and local label context.

A.3 Evaluation and NSA Scoring Protocol

Downstream model evaluation. All downstream models use stratified 60/20/20 train/validation/test splits, except on ogbn-arxiv, where we use OGB's official split. We report test micro-accuracy averaged over 10 random seeds. Validation data are used for early stopping and hyperparameter selection.

NSA scoring. Each NSA coordinate is obtained from the cross-validated macro-accuracy s of its fixed logistic-regression probe and normalized by the C-class chance level:

$$\widetilde{s} = \frac{s - 1/C}{1 - 1/C}.$$

Thus, $\widetilde{s} = 0$ corresponds to chance-level probe performance and $\widetilde{s} = 1$ to perfect prediction, independently of the number of classes. We use macro-accuracy for the probe because its chance baseline is $1/C$ even under class imbalance. The same scoring protocol is used for FSS, TSS, and LSS throughout the paper. The probe is multinomial logistic regression with ℓ_2 regularization at a fixed inverse strength $C_{\text{reg}} = 1.0$, solved with L-BFGS for at most 1000 iterations, with uniform class weights. These settings are held identical across all datasets and all three channels.

NSA scores are averaged over three annulus-sampling seeds. Unless explicitly stated otherwise, all NSA hyperparameters are fixed globally across datasets: projection dimension $r = 32$, $T = 4$ Weisfeiler–Lehman rounds, and annulus sample size $m_k = 64$.

A.4 Implementation Details

All GNN baselines use hidden width 64 and are trained for at most 500 epochs with validation-based early stopping of patience 50. We tune learning rate over $\{0.01, 0.005\}$, weight decay over $\{0,\ 5 \times 10^{-4},\ 5 \times 10^{-3}\}$, and dropout over $\{0.3,\ 0.5\}$. LP uses $\alpha \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99\}$. SGFormer uses three layers and four attention heads. GPS uses two layers with 16-dimensional random-walk positional encodings. Architecture-specific settings not listed here follow the standard implementations used in our code release.

A.5 Why NSA Uses Normalized Macro-Accuracy

NSA is intended to compare the accessibility of predictive signal across datasets with different numbers of classes and different class distributions. Raw micro-accuracy is poorly suited to this purpose: its baseline depends on class imbalance, so the same numerical accuracy can represent very different amounts of class-discriminative information across datasets. This issue is substantial in our benchmark suite, where the number of classes ranges from $C = 2$ to $C = 40$.

We therefore score each NSA probe using macro-accuracy (balanced accuracy), i.e., mean per-class recall. Its chance baseline is $1/C$, independent of class frequencies. We further normalize

$$\widetilde{s} = \frac{s - 1/C}{1 - 1/C},$$

where s is the cross-validated macro-accuracy of the probe. Thus, $\widetilde{s} = 0$ denotes chance-level class discrimination and $\widetilde{s} = 1$ denotes perfect discrimination; negative values indicate below-chance performance. This normalization places all datasets on the same chance-to-perfect scale.

This scoring choice applies only to the *NSA diagnostic*. Downstream node classification models are evaluated using the standard micro-accuracy protocol described in Appendix A.3.

Table 7 illustrates the effect on the 20 real benchmarks. The two protocols remain positively correlated ($r = 0.678$ for FSS, 0.493 for TSS, and 0.810 for LSS), but their absolute scales can differ substantially. The mean absolute difference is 0.211 for FSS, 0.262 for TSS, and 0.218 for LSS. The effect is especially visible on datasets with few classes or substantial imbalance. For example, Deezer-Europe has micro-accuracy probe scores around 0.56–0.57, which may superficially suggest moderate signal, whereas its chance-normalized NSA coordinates are only $(\text{FSS}, \text{TSS}, \text{LSS}) = (0.091, 0.010, 0.060)$. Genius is the sharpest case in the suite: its probes reach 0.80–0.81 under micro-accuracy, yet its chance-normalized coordinates are only $(0.037, 0.070, 0.086)$, because 80% of its nodes carry the majority label. Similarly, Actor's raw TSS of 0.252 becomes 0.004 after chance-relative scoring, correctly indicating essentially no recoverable structural-role signal.

The normalization can therefore change not only scale but also cross-dataset ranking, particularly for TSS. For example, Actor and Flickr have nearly identical micro-accuracy TSS values (0.252 and 0.258), while their normalized scores (0.004 and 0.162) clearly separate their structural signal levels. For a dataset atlas whose purpose is precisely cross-dataset comparison, this chance-relative interpretation is essential. All NSA scores reported in the paper use normalized macro-accuracy.

Table 7: **Effect of the NSA scoring protocol.** Normalized macro-accuracy and raw micro-accuracy of the three fixed probes on the 20 real benchmarks. Normalized macro-accuracy maps chance to 0 and perfect classification to 1 independently of class count and class imbalance. † denotes the filtered variant.

			Normalized macro-accuracy			Micro-accuracy		
Dataset	h_{node}	C	FSS	TSS	LSS	FSS	TSS	LSS
Roman-Empire	0.046	18	0.537	0.131	0.203	0.637	0.334	0.354
Texas	0.056	5	0.415	0.411	0.474	0.700	0.692	0.711
Wisconsin	0.099	5	0.337	0.139	0.150	0.692	0.497	0.537
Squirrel†	0.191	5	0.102	0.201	0.071	0.315	0.414	0.342
Cornell	0.200	5	0.298	0.097	0.105	0.517	0.383	0.381
Actor	0.202	5	0.120	0.004	0.007	0.319	0.252	0.253
Squirrel	0.215	5	0.286	0.200	0.125	0.429	0.360	0.299
Flickr	0.243	9	0.241	0.162	0.343	0.327	0.258	0.417
Chameleon†	0.244	5	0.199	0.234	0.140	0.389	0.422	0.354
Chameleon	0.247	5	0.511	0.321	0.282	0.610	0.462	0.431
Amazon-Ratings	0.376	5	0.076	0.024	0.149	0.397	0.371	0.454
BlogCatalog	0.391	6	0.466	0.288	0.594	0.558	0.409	0.665
Genius	0.509	2	0.037	0.070	0.086	0.805	0.812	0.804
Deezer-Europe	0.530	2	0.091	0.010	0.060	0.572	0.558	0.568
Ogbn-arxiv	0.635	40	0.413	0.059	0.481	0.641	0.324	0.702
CiteSeer	0.706	6	0.399	0.182	0.552	0.529	0.363	0.671
PubMed	0.792	3	0.678	0.315	0.711	0.794	0.573	0.824
Amazon-Computers	0.800	10	0.728	0.342	0.835	0.795	0.605	0.881
Cora	0.825	7	0.647	0.287	0.766	0.722	0.470	0.819
Amazon-Photo	0.850	8	0.823	0.577	0.876	0.867	0.726	0.910

A.6 Computational Cost

NSA is computed once per dataset before downstream model training. Table 8 reports wall-clock time for computing all three NSA coordinates on the 20 real benchmarks on 16 cores of a dual-socket AMD EPYC 9334 node (64 cores, 384 GB RAM). The pipeline is implemented entirely in NumPy and scikit-learn and requires no GPU, so it does not compete with downstream training for accelerator time. The reported times include neighborhood traversal, descriptor construction, and cross-validated fitting of the three fixed probes.

Table 8: Wall-clock time (seconds) for computing NSA scores on 20 real benchmarks, ordered by increasing runtime. All measurements were obtained on the compute node described above.

Dataset	*Texas*	*Cornell*	*Wisconsin*	*Chameleon†*	*Cora*	*CiteSeer*	*Chameleon*	*Squirrel†*	*Actor*	*PubMed*	*A. Ratings*	*A. Photo*	*D. Europe*	*R. Empire*	*Squirrel*	*BlogCatalog*	*Flickr*	*A. Computers*	*Ogbn-arxiv*	*Genius*
Time (s)	0.5	0.5	0.6	2.3	5.1	5.1	8.0	11.6	19.8	35.3	42.1	46.1	54.1	59.1	80.9	82.4	171.1	193.9	3,449.9	3,638.3

Runtime is small on the majority of benchmarks: the WebKB datasets complete in under one second, while medium-sized graphs such as Cora, Chameleon, and PubMed require a few seconds; the median runtime across the 20 benchmarks is 38.7 seconds. On these datasets, the one-time cost of computing NSA is clearly negligible relative to training even a single downstream architecture, let alone a full model search. Amazon-Computers, the largest of the mid-sized graphs, takes about 3.2 minutes. The cost becomes material only at OGB scale: Ogbn-arxiv (169,343 nodes) takes 57.5 minutes and Genius (421,961 nodes) 60.6 minutes, which is what motivates the subsampling study below.

Subsampling for large graphs. To evaluate subsampling as a way to reduce computational cost, we ran NSA on Ogbn-arxiv using 10k, 40k and 60k training nodes, drawn from the training split with its class proportions, alongside the full training split. The graph itself is never modified – all 2,315,598 edges are used for neighborhood construction in every row, so only descriptor construction and probe fitting scale down. The full computation takes 3,449.9 seconds (about 57.5 minutes). Subsampling reduces this to 1,137.3, 2,143.8 and 2,610.1 seconds, corresponding to 3.03×, 1.61× and 1.32× speedups. The saving is therefore real but modest: because the graph is left intact,

Table 9: NSA scores and computation time on Ogbn-arxiv under subsampling of the training nodes. Only the number of training nodes the probes are fitted and evaluated on is reduced; the graph is unchanged, so every row uses all 2,315,598 edges. Fractions are of the 90,941-node training split of OGB's time-based split. All four rows were measured on the same node with 16 CPU cores. Speedup is measured relative to the full computation.

Train nodes	Fraction	FSS	TSS	LSS	Time (s)	Speedup
10,000	11.0%	0.327	0.057	0.303	1,137.3	3.03×
40,000	44.0%	0.394	0.059	0.430	2,143.8	1.61×
60,000	66.0%	0.408	0.059	0.458	2,610.1	1.32×
90,941	100.00%	0.413	0.059	0.481	3,449.9	1.00×

neighborhood expansion costs the same in every row, so a $9\times$ reduction in the number of probed nodes buys only a $3.03\times$ speedup.

Table 9 illustrates the associated accuracy–cost trade-off. As the sample size increases, FSS and LSS move closer to their full-graph values. At 40k training nodes (44.0% of the training split), their absolute differences from the full scores are 0.019 and 0.051, while runtime is reduced by 37.9%. At 60k training nodes (66.0%), these differences decrease to 0.005 and 0.023, with a 24.3% runtime reduction. TSS is essentially flat across all four rows (0.057–0.059), which is what one expects when the graph is untouched; LSS degrades fastest because shrinking the node set also shrinks the set of labels visible to its descriptor. Thus, node subsampling offers a practical way to reduce NSA's computational cost when approximate scores are sufficient. These results demonstrate the potential on one large graph; the sample size needed for a desired accuracy remains dataset-dependent.

B Extended Analysis on Synthetic Benchmarks

The controlled synthetic study tests whether NSA behaves as a measurement instrument when the underlying source of predictive signal is known. We evaluate three properties: (i) *calibration*, whether each coordinate responds to changes in its corresponding signal; (ii) *compositionality*, whether mixed-signal problems occupy intermediate positions in NSA space; and (iii) *scope*, whether NSA correctly exposes cases in which predictive information lies outside its three marginal channels.

The suite contains 24 synthetic datasets: eight corner cases spanning the $2 \times 2 \times 2$ combinations of feature (F), topology (T), and label-context/smoothness (L) signal, together with 16 intermediate-strength datasets. All datasets use $N = 1500$ nodes, $C = 3$ classes, feature dimension $d = 16$, and the same 60/20/20 train/validation/test protocol used for the real benchmarks. Results are averaged over 10 random seeds and reported in Table 10.

B.1 Construction and Signal Controls

The three channels are controlled through distinct mechanisms.

Feature signal (F). When F is active, node features are sampled from separated class-conditional Gaussian distributions, making labels recoverable from feature geometry. When F is suppressed, features are replaced by class-independent noise.

Topology signal (T). When T is active, classes occupy graph blocks with distinct structural signatures under a degree-corrected stochastic block model (DC-SBM). The resulting classes differ through structural quantities such as degree, expansion behavior, and Weisfeiler–Lehman role statistics. When T is suppressed, structural characteristics are matched across classes so that unlabeled topology alone is not predictive.

Label-context signal (L). When L is active, the graph is assortative with respect to class labels, so observed labels in local neighborhoods are informative about the target node. When L is suppressed, edge formation is approximately label-independent.

T–L coupling. The three controls are not perfectly orthogonal. In particular, making structural roles class-dependent also induces label dependence in local graph context, so strong T signal can elevate LSS. The converse is much weaker: the L construction degree-matches the blocks and can therefore produce high LSS without high TSS. This asymmetry is visible in the pure T and L corners: the T corner has $(\text{TSS}, \text{LSS}) = (0.711, 0.786)$, whereas the L corner has $(\text{TSS}, \text{LSS}) = (0.142, 0.773)$. We therefore do not interpret the synthetic cube as requiring perfect statistical independence of the three measured coordinates; rather, NSA should report the signal structure actually induced by the graph. The same coupling reappears, in a different construction, in the `structural_role` regime of the architecture-triage benchmark (Appendix D).

B.2 Corner and Boundary Cases

The corner datasets provide the clearest controlled tests.

Pure channels. The F corner cleanly isolates feature signal:

$$(\text{FSS}, \text{TSS}, \text{LSS}) = (0.681, -0.012, 0.001).$$

MLP reaches 86.70%, while label propagation remains near chance. The L corner similarly produces high label-context signal,

$$(\text{FSS}, \text{TSS}, \text{LSS}) = (0.251, 0.142, 0.773),$$

and LP reaches 89.43%. The T corner reflects the T–L coupling discussed above,

$$(\text{FSS}, \text{TSS}, \text{LSS}) = (0.257, 0.711, 0.786),$$

while graph-aware models substantially outperform MLP and LP. Thus, the coordinates respond to the intended signal manipulations without requiring that every synthetic mechanism isolate exactly one axis.

Mixed corners. The FT, FL, TL, and FTL datasets activate multiple sources simultaneously. Their NSA coordinates reflect the corresponding combinations: F-active datasets have elevated FSS, L-active datasets have elevated LSS, and T-active datasets exhibit elevated TSS together with the expected T–L coupling. Downstream behavior is correspondingly less tied to a single interface, illustrating why NSA should be interpreted as a signal profile rather than a discrete dataset taxonomy.

Boundary case: Node Twins. Dataset C is deliberately constructed so that labels are encoded through a *joint* relationship between feature type and structural role rather than through any individual marginal channel. NSA correctly reports essentially no marginal signal,

$$(\text{FSS}, \text{TSS}, \text{LSS}) = (-0.016, -0.014, -0.018),$$

yet both graph transformers in our baseline set exploit it: GPS reaches 87.95% and SGFormer reaches 53.73%, against a 33.3% chance level. That two architecturally distinct models relying on global or joint graph-feature processing both clear chance is stronger evidence for a genuine scope boundary than either result alone. This example therefore marks the limit of NSA's marginal measurement: low NSA coordinates imply little signal in the three measured channels, not the impossibility of prediction by a richer model exploiting cross-channel interactions.

B.3 Intermediate Sweep and Full Results

The remaining 16 datasets introduce intermediate-strength controls: **Fm** reduces class separation in feature space, **Tm** reduces structural-role contrast, and **Lm** reduces label assortativity. Mixed datasets combine these medium and full-strength controls; for example, FmT has medium feature signal and full topology signal, while FmTmLm places all three mechanisms at intermediate strength. Exact generator parameters are provided in the accompanying code.

The single-channel sweeps show the expected calibration behavior. Weakening F from F to Fm reduces FSS from 0.681 to 0.331, while TSS and LSS remain essentially unchanged. Weakening L from L to Lm reduces LSS from 0.773 to 0.341. Weakening T from T to Tm reduces TSS from 0.711 to 0.254; LSS also decreases from 0.786 to 0.222, reflecting the known T–L coupling. The

Table 10: **Synthetic Experiments.** NSA scores and model performance on the 24 synthetic datasets ($N = 1500$, $C = 3$, mean over 10 seeds; chance = 33.3%). The first 8 rows are the corner cases of the F/T/L cube; the last 16 rows are intermediate cases. NSA scores are normalized macro-accuracy (0∼chance, 1∼perfect). **Best**, **second** and **third** performances are highlighted

Dataset	FSS	TSS	LSS	MLP	LP	APPNP	GCN	SAGE	H2GCN	SGC	GPS	SGFormer
F	0.681	-0.012	0.001	**86.70±1.38**	33.99±2.24	78.48±1.53	35.68±3.39	56.83±2.08	62.61±3.88	64.22±2.50	**85.54±1.44**	**85.87±1.00**
T	0.257	0.711	0.786	31.77±2.34	33.37±0.10	33.70±2.75	**68.10±4.04**	52.53±3.37	65.90±3.52	36.50±2.77	**74.97±2.54**	**74.93±2.99**
L	0.251	0.142	0.773	31.77±2.34	**89.43±1.41**	50.70±2.76	62.03±2.09	55.17±1.72	65.93±2.08	49.23±1.67	**71.93±2.73**	**79.37±1.65**
FT	0.864	0.711	0.786	**85.57±1.71**	33.37±0.10	76.50±1.67	70.40±3.05	74.67±2.65	82.27±2.25	83.40±2.53	**93.83±0.93**	**94.60±1.45**
FL	0.899	0.142	0.773	85.57±1.71	89.43±1.41	**95.83±1.19**	72.03±1.37	78.03±1.76	83.83±1.80	94.67±1.02	**94.97±1.62**	**95.37±0.85**
TL	0.520	0.928	0.911	31.77±2.34	79.60±1.79	55.13±2.30	70.97±2.55	62.23±1.48	**82.43±2.66**	48.17±1.34	**86.37±1.73**	**85.47±1.12**
FTL	0.927	0.928	0.911	85.57±1.71	79.60±1.79	**95.93±0.99**	81.53±2.10	78.33±1.34	90.00±1.29	94.60±1.64	**96.13±0.90**	**97.17±0.83**
C	-0.016	-0.014	-0.018	32.67±2.57	33.14±2.58	**38.61±2.90**	33.80±2.02	33.14±2.11	32.74±2.16	38.15±2.70	**87.95±3.13**	**53.73±2.82**
Fm	0.331	-0.012	0.001	**60.92±1.63**	33.99±2.24	55.97±1.22	32.38±1.68	31.91±2.05	33.47±2.68	51.39±2.00	**61.06±1.44**	**60.00±2.31**
Lm	0.080	-0.036	0.341	31.77±2.34	**57.77±2.04**	**41.07±2.22**	40.57±2.40	37.07±3.01	39.10±3.96	39.67±3.52	**43.33±3.13**	42.77±2.84
Tm	0.133	0.254	0.222	31.77±2.34	36.50±2.25	37.60±2.41	**43.03±1.62**	37.00±2.97	36.70±1.61	36.13±2.11	**42.67±3.48**	**42.17±2.04**
FLm	0.711	-0.036	0.341	**85.57±1.71**	57.77±2.04	**87.57±1.71**	50.07±1.44	67.43±2.43	69.13±2.33	77.20±2.15	**87.23±1.86**	**87.57±2.10**
FmL	0.687	0.142	0.773	60.47±2.11	**89.43±1.41**	81.03±1.36	60.27±2.95	57.97±2.00	67.63±2.58	**85.40±1.23**	85.37±1.64	**88.50±1.88**
FmLm	0.378	-0.036	0.341	60.47±2.11	57.77±2.04	**66.97±1.91**	43.07±1.51	39.83±3.28	38.57±4.14	62.07±1.80	**66.13±1.82**	**65.00±1.51**
FmT	0.466	0.483	0.433	**60.47±2.11**	40.07±1.95	60.27±2.95	53.47±3.46	42.73±1.90	52.73±2.34	55.03±2.61	**67.13±2.93**	**67.37±2.96**
FmTL	0.753	0.928	0.911	60.47±2.11	79.60±1.79	81.53±0.91	74.27±2.59	65.40±2.32	84.37±1.97	**85.07±1.72**	**92.03±1.80**	**92.43±1.97**
FmTm	0.359	0.254	0.222	**60.47±2.11**	36.50±2.25	58.13±3.28	45.53±3.51	37.10±3.16	38.83±4.08	52.93±3.69	**61.87±2.44**	**62.27±1.92**
FmTmLm	0.453	0.434	0.531	60.47±2.11	60.30±1.22	**67.70±2.33**	51.93±2.63	42.57±2.71	52.03±4.11	64.23±1.25	**70.00±2.33**	**71.67±1.67**
FTLm	0.817	0.673	0.655	85.57±1.71	59.13±1.31	**88.77±1.67**	66.07±3.74	68.23±2.93	79.33±2.71	82.43±1.29	**90.87±1.26**	**91.23±1.16**
FTm	0.688	0.254	0.222	**85.57±1.71**	36.50±2.25	79.90±2.38	47.83±2.32	64.50±2.60	66.97±3.26	66.43±2.67	**83.93±2.14**	**84.70±1.95**
FTmL	0.891	0.837	0.848	85.57±1.71	84.43±1.47	**95.50±1.39**	76.40±2.28	77.60±3.19	85.33±1.93	94.33±0.98	**94.63±1.60**	**96.47±1.03**
TLm	0.400	0.673	0.655	31.77±2.34	59.13±1.31	45.97±2.13	58.77±1.67	46.83±2.67	**64.77±1.52**	40.17±2.81	**66.43±2.52**	**62.87±2.20**
TmL	0.374	0.837	0.848	31.77±2.34	**84.43±1.47**	51.50±2.30	67.23±3.72	57.83±2.67	74.90±1.84	48.53±1.56	**84.00±1.90**	**76.90±2.38**
TmLm	0.249	0.434	0.531	31.77±2.34	**60.30±1.22**	43.77±2.74	50.17±3.52	42.63±2.99	51.83±2.41	39.83±2.57	**55.27±3.27**	**53.00±3.18**

corresponding downstream interfaces change in the same direction: MLP accuracy decreases from 86.70% to 60.92% under F→Fm, while LP decreases from 89.43% to 57.77% under L→Lm.

Mixed intermediate datasets also populate graded regions of NSA space rather than collapsing onto the corners. For example, FmTmLm produces

$$(\text{FSS}, \text{TSS}, \text{LSS}) = (0.453, 0.434, 0.531),$$

placing it in the interior of the three-dimensional profile. Across the full suite, these results support the intended interpretation of NSA as a continuous description of accessible signal rather than a collection of binary regime labels.

B.4 Guidance Rule

NSA converts a dataset's (FSS, TSS, LSS) profile into a coarse interface-level recommendation using a fixed activity threshold $\delta = 0.05$ and margin threshold $\varepsilon = 0.10$, applied identically to every dataset in the paper, real or synthetic. A coordinate is *active* if it exceeds δ. If no coordinate is active, the dataset is a low-signal (*Desert*) zone. Otherwise, the dataset is assigned a *singleton* zone when its leading coordinate exceeds the next-highest active coordinate by at least ε, and a *mixed* zone otherwise. Table 11 lists the resulting eight zones and their recommended first step.

Two boundary conventions make the rule fully deterministic. First, the margin comparison uses $\geq \varepsilon$, so a leading coordinate exactly at the margin resolves to the corresponding singleton zone; this places borderline cases such as Squirrel† (TSS − FSS = 0.099) in the leading (T) zone rather than leaving them ambiguous. Second, the Desert zone captures datasets in which every coordinate is close to chance: a dataset whose *leading* coordinate only marginally exceeds δ, with the remaining two at or below it, carries essentially no accessible marginal signal and is read as low-signal. Actor $(0.120, 0.004, 0.007)$ is the clearest such case, and we describe it as a signal desert throughout for this reason.

The rule is intentionally simple. F-dominant datasets prioritize feature-centric models, L-dominant datasets prioritize diffusion models, and T-dominant datasets prioritize topology-aware GNNs; mixed zones prioritize the corresponding combination of interfaces. This is the rule referenced wherever a dataset is described by its NSA regime (e.g., Table 13).

Table 11: **NSA guidance rule.** A zone is assigned using activity threshold $\delta = 0.05$ and margin threshold $\varepsilon = 0.10$, translating atlas position into a first-pass model-search recommendation.

Zone	Signal pattern	Recommended first step
F	FSS dominant	Include feature-centric models early
L	LSS dominant	Include diffusion models early
T	TSS dominant	Include topology-aware GNNs early
FL	FSS, LSS active	Try feature-centric and diffusion first
FT	FSS, TSS active	Try feature-centric and topology-aware first
TL	TSS, LSS active	Try topology-aware and diffusion first
FTL	all channels active	Try graph-aware families first; broaden if needed
Desert	no active channel	Broaden search; no narrow recommendation

Table 12: **Model performance across datasets.** Node classification performance on 20 real benchmarks, reported as mean test accuracy (%)± standard deviation. Blue, magenta, and teal indicate the best, second-best, and third-best results, respectively. OOM denotes out-of-memory failure.

Dataset	MLP	LP	APPNP	GCN	SAGE	H2GCN	SGC	GPS	SGFormer
roman-empire	66.12±0.65	8.08±0.30	56.33±0.57	36.56±0.62	66.30±0.53	**70.84±0.49**	41.75±0.47	**82.75±0.41**	**80.07±0.51**
texas	78.78±4.09	19.27±2.98	51.71±3.05	55.37±3.09	74.15±5.02	**80.00±4.34**	50.24±2.93	**80.73±4.00**	**80.49±4.22**
wisconsin	86.42±2.50	17.92±4.72	51.32±6.01	49.43±3.56	**82.08±4.94**	81.70±5.20	46.04±3.50	**82.26±5.01**	**86.42±2.64**
squirrel-fltr	35.41±1.70	32.94±1.65	32.74±2.21	**37.93±2.05**	33.25±1.15	32.96±1.82	**38.02±1.93**	**39.69±1.18**	36.24±2.22
cornell	**75.75±4.88**	23.25±2.97	47.50±3.71	42.25±5.30	71.25±5.15	69.25±3.88	42.25±2.84	**75.25±2.36**	**75.50±5.45**
actor	**36.73±0.94**	18.36±1.27	29.00±1.13	29.19±0.93	**37.15±1.05**	36.70±1.01	28.73±1.60	**37.57±1.13**	36.51±1.18
squirrel	34.57±1.56	16.57±0.98	29.32±1.90	28.62±1.24	33.11±1.00	**43.50±1.85**	27.39±1.60	**35.70±1.49**	**35.51±1.58**
flickr	91.40±0.87	41.34±1.55	92.17±0.70	70.81±1.09	**92.33±0.56**	**92.70±0.58**	67.84±0.80	90.52±0.87	**92.32±0.68**
chameleon-fltr	**39.18±4.62**	33.28±2.22	38.42±1.94	36.17±2.26	35.85±2.58	36.23±2.28	36.23±3.67	**41.15±3.32**	**39.34±3.67**
chameleon	51.33±2.25	21.15±2.07	43.64±2.30	42.88±2.84	49.83±2.50	**63.18±2.14**	38.58±2.95	**52.22±2.04**	**52.88±2.09**
amazon-ratings	44.27±0.90	**46.98±0.47**	46.13±0.62	43.37±0.78	43.20±0.88	45.49±0.56	41.94±0.79	**48.81±0.90**	**50.88±0.46**
blogcatalog	93.18±0.90	65.34±1.00	95.57±0.71	77.93±1.01	95.56±0.61	**96.38±0.33**	78.19±1.07	**95.89±0.58**	**96.30±0.60**
genius	82.55±0.10	80.36±0.06	81.38±0.39	82.43±0.36	**84.49±0.13**	**84.86±0.09**	81.91±0.32	OOM	**85.55±0.14**
deezer-europe	**67.47±0.44**	56.51±0.29	65.60±0.59	61.99±0.39	66.76±0.33	**67.33±0.50**	62.97±0.48	66.89±0.66	**67.51±0.48**
ogbn-arxiv	52.48±0.27	**67.62±0.00**	**68.48±0.27**	59.86±0.37	60.75±0.43	63.76±0.29	65.84±0.06	OOM	**70.47±0.23**
citeseer	72.83±0.91	70.64±1.25	**77.21±0.79**	76.86±1.17	**77.45±1.28**	**77.18±1.10**	76.81±1.32	76.35±1.18	76.32±1.06
pubmed	87.83±0.53	83.20±0.57	88.27±0.52	87.68±0.52	88.93±0.53	**89.63±0.32**	84.78±0.63	**89.34±0.43**	**89.91±0.54**
amazon-comp	85.34±0.50	89.35±0.57	**90.70±0.42**	88.57±0.51	88.13±5.31	90.31±0.48	89.67±0.44	**90.91±0.64**	**91.54±0.63**
cora	74.74±1.89	84.87±0.83	**88.81±0.81**	87.01±1.00	87.50±1.12	**87.94±1.52**	**87.72±1.38**	84.60±1.11	85.42±1.48
amazon-photo	92.40±0.78	91.64±0.66	**95.19±0.60**	93.54±0.45	**95.21±0.66**	95.09±0.61	93.76±0.48	94.68±0.69	**95.69±0.56**

C Extended Analysis on Real Benchmarks

This section provides the full model results and additional analyses supporting the interpretation of NSA on the 20 real benchmarks. We focus on three questions: how the datasets populate NSA space, what information NSA provides beyond scalar homophily, and how the individual coordinates relate to the model interfaces they are intended to characterize.

C.1 Full Results and the Real-Data Atlas

Table 12 reports mean test accuracy over 10 random seeds for all nine downstream models. Together with the NSA coordinates in Table 3, these results illustrate that graphs with similar homophily can present substantially different learning problems.

Figure 1 visualizes the 20 datasets in $(\text{FSS}, \text{TSS}, \text{LSS})$ space. The benchmarks occupy a broad region rather than collapsing onto a single homophily-like direction. The clearest global pattern is that high-homophily datasets concentrate toward larger LSS, consistent with homophily primarily describing local label alignment. By contrast, low-homophily datasets are widely dispersed along the other coordinates, particularly FSS. Thus, low homophily does not imply weak accessible signal.

Several datasets illustrate these distinctions. Roman-Empire and Texas have very low node homophily but substantial feature accessibility, showing that a heterophilic graph need not be signal-poor. Actor exhibits a different type of difficulty: all three coordinates are small ($\text{FSS} = 0.120$, $\text{TSS} = 0.004$, $\text{LSS} = 0.007$), and performance is comparatively compressed across model families. Deezer-Europe similarly has uniformly low NSA coordinates $(0.091, 0.010, 0.060)$, indicating that none of the three marginal channels provides a strong signal.

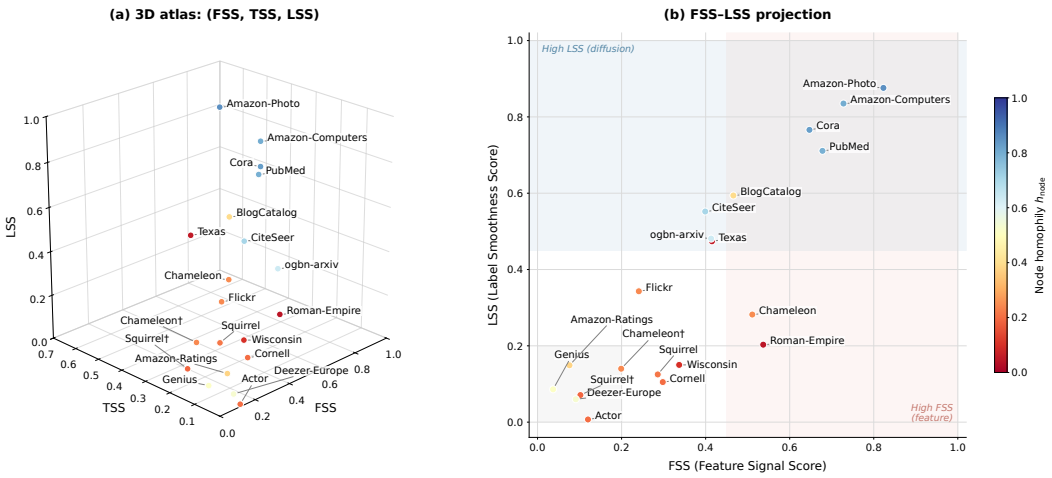

Figure 1: **Node Signal Atlas for 20 real benchmarks. (a)** 3D scatter of all 20 datasets in (FSS, TSS, LSS) space. **(b)** FSS–LSS projection, with light shading indicating the high-LSS zone (diffusion-dominant), high-FSS zone (feature-dominant), and low-signal corner. Color encodes node homophily h_{node} (blue = high, red = low). High-homophily datasets concentrate in the high-LSS region, confirming that homophily tracks the smoothness channel. Low-homophily datasets scatter widely along the FSS axis, showing that heterophilic does not imply signal-poor. Actor lies near the low-signal corner with all three scores close to zero, consistent with its uniformly low model accuracies across all families.

The raw/filtered benchmark pairs provide another useful check. Chameleon's FSS drops from 0.511 to 0.199 after filtering, while its node homophily changes only slightly. Squirrel shows a similar reduction in FSS, from 0.286 to 0.102, while TSS remains essentially unchanged (0.200 vs. 0.201). NSA therefore exposes changes in signal composition that are largely invisible to a scalar homophily score.

C.2 Beyond Homophily: Similar Homophily, Different Problems

Node homophily measures a specific aspect of the learning problem: the tendency of adjacent nodes to share labels. In NSA terms, this is most closely related to local label context and therefore to LSS. It does not directly measure feature discriminability or unlabeled structural-role information. Consequently, datasets with nearly identical homophily can occupy very different locations in NSA space.

Table 13 gives four representative pairs with $\Delta h_{\text{node}} \leq 0.010$. Despite nearly identical homophily, the paired datasets differ substantially in their NSA profiles and empirical difficulty.

Cornell vs. Actor. Cornell and Actor have nearly identical node homophily (0.200 vs. 0.202), but their signal profiles are very different. Cornell has appreciable feature signal (FSS $= 0.298$) and some label-context signal (LSS $= 0.105$), whereas Actor is close to the low-signal corner across all three coordinates. Their best model accuracies differ accordingly, 75.75% versus 37.57%. Homophily places the two graphs in essentially the same regime; NSA distinguishes one as having accessible feature signal and the other as having little signal in any measured marginal channel.

Cornell vs. Squirrel-filtered. This pair isolates a different comparison: two datasets with a dominant signal in *different* channels rather than one signal-rich and one signal-poor. Cornell ($h_{\text{node}} = 0.200$) is feature-dominant (FSS $= 0.298$), while Squirrel-filtered ($h_{\text{node}} = 0.191$) has its largest coordinate in TSS (0.201, versus FSS $= 0.102$), placing it in a mixed/topology-leaning region rather than Cornell's feature-dominant one. Their best accuracies differ substantially (75.75% vs. 39.69%), and critically the two datasets are best served by different diagnostic signals – not merely different amounts of the same one. Homophily alone ($\Delta h_{\text{node}} = 0.009$) gives no indication of this difference.

Table 13: Dataset pairs with similar h_{node} ($\Delta h_{\text{node}} \leq 0.010$) but different NSA profiles and empirical outcomes. **NSA Regime** summarizes the signal pattern: *F* = feature-dominant, *L* = smoothness-dominant, *mixed* = no dominant channel. **Best Acc.** = highest accuracy across all nine baselines. † = filtered variant.

Dataset	h_{node}	FSS	TSS	LSS	NSA Regime	Best Family	Best Acc.
Flickr	0.243	0.241	0.162	0.343	L	T	92.70
Chameleon†	0.244	0.199	0.234	0.140	mixed	GT	41.15
Cornell	0.200	0.298	0.097	0.105	F	F	75.75
Actor	0.202	0.120	0.004	0.007	low	GT	37.57
Cornell	0.200	0.298	0.097	0.105	F	F	75.75
Squirrel†	0.191	0.102	0.201	0.071	T	GT	39.69
Chameleon	0.247	0.511	0.321	0.282	F	T	63.18
Chameleon†	0.244	0.199	0.234	0.140	mixed	GT	41.15

Flickr vs. Chameleon-filtered. These datasets have almost identical homophily (0.243 vs. 0.244), yet Flickr has substantially larger LSS (0.343 vs. 0.140) and achieves a much higher performance ceiling (92.70% vs. 41.15%). Their similar homophily therefore conceals substantial differences in both signal composition and empirical difficulty.

Chameleon before and after filtering. Raw and filtered Chameleon have node homophily 0.247 and 0.244, respectively, but filtering reduces FSS from 0.511 to 0.199 and lowers the best-model accuracy from 63.18% to 41.15%. This pair is particularly informative because the graph source is closely related while the accessible signal changes substantially. The FSS coordinate detects this change directly, whereas homophily is nearly unchanged. We note that raw Chameleon's best performing model (H2GCN, 63.18%) belongs to the topology-aware family even though FSS is its dominant coordinate; a high FSS indicates that feature-centric models are worth including in a shortlist, not that they must win outright, and Chameleon's mixed FSS/TSS profile (0.511/0.321) is consistent with a topology-aware model remaining competitive.

These cases illustrate the intended role of NSA. Homophily remains informative about label alignment, but the three-coordinate representation separates learning problems that a single scalar can place in nearly the same regime.

C.3 TSS as a Conditional Signal

Among the three coordinates, TSS has the most conditional interpretation. It measures whether class labels are recoverable from *unlabeled structural-role descriptors*, and is not intended as a universal predictor of model accuracy: structural information helps only when the downstream architecture can exploit it and when its contribution is not dominated by stronger feature or label-context signal. We examine this conditionality from two angles, a targeted gain analysis and descriptive interface-level correlations.

Topology-aware gain over feature-only. The natural targeted quantity is the accuracy gained by adding a topology-aware interface over the feature-only baseline:

$$\Delta_{\text{GNN}} = \max_{\mathcal{M} \in \{\text{GCN,SAGE,H2GCN,SGC}\}} \text{Acc}(\mathcal{M}) - \text{Acc}(\text{MLP}).$$

Split at TSS $= 0.20$, the two halves of the suite behave differently. Among the ten datasets with TSS ≥ 0.20, a topology-aware GNN improves over MLP on nine, with a mean gain of 4.8 percentage points; among the ten with TSS < 0.20, it improves on only seven, with a mean gain of 1.4 points, and that smaller figure is itself carried by a single outlier (OGBN-ARXIV). The contrast is therefore clearest in how *reliably* structure helps rather than in the average magnitude: elevated TSS flags the datasets on which adding graph structure to a feature-only model is likely to pay off. TSS is thus weaker than FSS or LSS as a global correlate of raw accuracy, but informative for the narrower question it is designed to answer, whether unlabeled structure carries class-relevant information that a topology-aware model can use.

Table 14: Pearson r and Spearman ρ correlations between NSA scores and interface-level model performance across real benchmarks. $^{**}\ p < 0.01$, $^{*}\ p < 0.05$.

	acc_F		acc_L		acc_G	
Score	r	ρ	r	ρ	r	ρ
FSS	0.510^{*}	0.472^{*}	0.555^{*}	0.517^{*}	0.622^{**}	0.611^{**}
TSS	0.408	0.389	0.388	0.358	0.423	0.457^{*}
LSS	0.595^{**}	0.612^{**}	0.771^{**}	0.744^{**}	0.674^{**}	0.746^{**}

Interface-level correlations. As a descriptive sanity check, we relate the three coordinates to performance at the level of broad model interfaces. We define

$$\mathrm{Acc}_F(D) = \mathrm{Acc}(\mathrm{MLP}), \qquad \mathrm{Acc}_L(D) = \max\{\mathrm{Acc}(\mathrm{LP}), \mathrm{Acc}(\mathrm{APPNP})\},$$

$$\mathrm{Acc}_G(D) = \max\{\mathrm{Acc}(\mathrm{GCN}), \mathrm{Acc}(\mathrm{SAGE}), \mathrm{Acc}(\mathrm{H2GCN}), \mathrm{Acc}(\mathrm{SGC}), \mathrm{Acc}(\mathrm{GPS}), \mathrm{Acc}(\mathrm{SGFormer})\}.$$

Across the 20 real benchmarks, two patterns are clear. FSS and LSS correlate strongly with performance across several interfaces: FSS between 0.51 and 0.62, reflecting the broad value of discriminative node features even for architectures that subsequently process the graph, and LSS more strongly still, between 0.60 and 0.77, consistent with local label context benefiting many propagation-based architectures rather than only explicit label propagation. TSS, by contrast, correlates only weakly with absolute accuracy ($r < 0.50$ at every interface). This does not contradict its intended role: TSS measures the *availability* of structural-role information, whereas overall benchmark accuracy is typically dominated by features or smoothness. The gain analysis and the correlations make the same point from two directions, that TSS is best read as a complementary structural axis, informative for the conditional question of when topology helps, rather than as a standalone measure of task difficulty. Because the sample of 20 benchmarks is modest, we treat these correlations as descriptive supporting evidence rather than as a model-selection result.

Overall, the real-benchmark analysis supports the central distinction behind NSA: homophily primarily describes label–adjacency alignment, while feature accessibility and unlabeled structural information vary substantially even among graphs with similar homophily. The three coordinates therefore provide a richer description of the learning problem without implying that any individual coordinate should uniquely determine the best architecture.

C.4 Comparison with Related Characterization Approaches

A growing body of work has shown that standard homophily is not sufficient to explain node-classification behavior. Several recent approaches refine homophily, measure when graph aggregation helps, or combine multiple graph statistics into stronger predictive metrics, while a separate line of work has scrutinized how such metrics and the models motivated by them are evaluated in the first place. NSA shares this motivation, but targets a different object than a metric-construction approach: it is designed as a multi-channel, interface-aligned diagnostic of accessible signal, computed before any downstream model is trained.

Adjusted homophily and label informativeness. Platonov et al. (Platonov et al., 2023a) show that standard edge and node homophily can be difficult to compare across datasets with different class counts or class imbalance, since their baselines depend on the label distribution. They propose adjusted homophily, h_{adj}, equivalent to the assortativity coefficient, as a more comparable measure of edge-label agreement. They also introduce label informativeness (LI), a normalized mutual-information measure between the labels of adjacent nodes, to distinguish informative heterophily from noisy neighbor labels.

NSA shares the same starting point: homophily alone is an incomplete description of node-classification signal. The difference is that h_{adj} and LI remain label-neighborhood summaries. They measure how labels interact across edges, but do not directly ask whether node features are discriminative or whether unlabeled structural roles are class-informative. NSA separates these effects into three axes. LSS captures the label-smoothness signal most closely related to homophily-style quantities, FSS captures feature accessibility, and TSS captures structural-role signal through unlabeled

ego-structure descriptors. Thus, NSA refines the homophily view not by replacing it with another scalar, but by embedding each dataset into a three-channel signal profile.

Node distinguishability and classifier-based diagnostics. Luan et al. (Luan et al., 2024b) study when graph neural networks help from a node-separability perspective. In a contextual stochastic block model, they derive distinguishability metrics that characterize how graph filtering affects intra-class and inter-class separation, including the "mid-homophily pitfall" where low-pass filtering can make nodes least distinguishable at intermediate homophily. They also propose a practical classifier-based performance metric (CPM), which compares the accuracy of a fixed classifier on raw features versus graph-aggregated features and uses this comparison to decide whether graph aggregation is likely to help over a coupled MLP baseline.

NSA is complementary to this viewpoint. CPM directly targets a binary question: whether graph aggregation improves a specific paired model interface, such as GCN versus MLP or SGC versus MLP. NSA instead asks a descriptive question: which sources of signal are accessible in the dataset before training a downstream GNN? Its output is therefore a continuous three-dimensional profile rather than a single decision or p-value. This distinction matters because graph-aware improvement can arise for different reasons. A dataset may benefit from propagation because labels are smooth, from structural modeling because unlabeled roles are informative, or from features alone because node attributes are already discriminative. NSA keeps these possibilities separate through FSS, TSS, and LSS. It also trains no downstream neural model; the only fitted models are fixed logistic-regression probes used uniformly to score the three channels.

Disentangled graph homophily and Tri-Hom. Zheng et al. (Zheng et al., 2024) argue that graph characterization should account for labels, structure, and features, rather than relying only on edge-label agreement. They propose three homophily notions: label homophily, structural homophily, and feature homophily, and combine them through a CSBM-3H-inspired scalar metric, Tri-Hom, designed to better correlate with GNN accuracy. This is close in spirit to NSA in that both approaches emphasize signal composition rather than a single homophily axis.

The main difference is the measurement philosophy. Tri-Hom constructs a composite scalar intended to improve prediction of GNN performance. NSA keeps the channels separate and independently interpretable. This separation is deliberate: FSS, TSS, and LSS are aligned with feature-centric, topology-aware, and diffusion-based model interfaces, respectively. As a result, NSA can indicate not only that a dataset is likely to be easy or hard, but also which kind of model family is plausible and why. Another difference is that NSA's topology descriptors are purely unlabeled structural summaries, including degree, core number, egonet density, expansion statistics, and WL role entropy. The resulting TSS score is still evaluated by a supervised probe, as are all NSA scores, but the structural representation being probed does not use labels or node features. This makes TSS a direct test of whether unlabeled graph structure alone carries class-relevant information.

Evaluation rigor and dataset taxonomies for heterophily. A related but distinct concern is not how a characterization metric is *constructed*, but how metrics and the models they motivate are *evaluated*. Luan et al. (Luan et al., 2024c) identify three pitfalls that have distorted the field's picture of heterophily-specific GNNs and homophily metrics – insufficient hyperparameter tuning of baselines, insufficient evaluation on genuinely challenging datasets, and, most relevant here, the absence of a quantitative benchmark for testing homophily metrics against known ground truth. After retuning baselines on 27 widely used datasets, they classify each as *malignant* (genuinely hard for standard GNNs even after tuning), *benign* (heterophilic by homophily score but easy in practice), or *ambiguous*.

This taxonomy and NSA's atlas ask related but distinct questions about the same underlying problem. Their taxonomy is discrete and post-hoc: it labels a dataset after training and tuning models on it. NSA's position in $(\text{FSS}, \text{TSS}, \text{LSS})$ space is continuous and a priori: it is computed before any downstream model is trained, and indicates not just whether a dataset is hard but where its recoverable signal actually lives. The Chameleon / Chameleon† pair discussed above (§C.2) is a concrete point of contact: filtering the leaked structural identifiers that made raw Chameleon look easier than it is would plausibly move that dataset from something close to their benign category toward their malignant one, and NSA locates the same shift directly, as a drop in FSS from 0.511 to 0.199 with node homophily nearly unchanged. Where their taxonomy would report that the dataset became

harder, NSA additionally reports *why*: the feature-accessible signal that made the raw graph artificially easy is no longer there. More broadly, their emphasis on validating metrics against synthetic data with known ground truth, rather than real-benchmark correlation alone, mirrors the design principle behind our own synthetic suite (Appendix B): a dataset-characterization score cannot be trusted from its correlation with real-benchmark accuracy alone, since real benchmarks confound many effects at once, and must also be validated where the signal source is known and can be manipulated in isolation. We arrived at this principle independently and for a different object – they validate whether homophily-style metrics track true dataset difficulty; we validate whether FSS, TSS, and LSS track true signal source – but the shared conclusion, that synthetic ground-truth validation is not optional for a characterization metric, reinforces the case for both.

Summary. Table 15 summarizes the main distinctions among the three characterization metrics discussed above (h_{adj}+LI, CPM, and Tri-Hom); the evaluation-rigor critique of Luan et al. (Luan et al., 2024c) is omitted from the table since it targets evaluation protocol rather than proposing a characterization metric of the same kind. Across all four, the central difference from NSA is that prior approaches primarily refine scalar homophily-style characterization, predict whether graph-aware models will outperform simpler baselines, or correct how such comparisons are evaluated. NSA instead provides a lightweight pre-training signal profile. It does not replace validation or model selection; rather, it explains which signal channels are present and therefore which model interfaces are worth validating first.

Table 15: Comparison of NSA with related graph characterization frameworks. "Pre-training" means the diagnostic can be computed before training a downstream GNN. "Multi-channel" means the method returns multiple independently interpretable scores rather than a single scalar. "Unlabeled structure" means that the structural descriptors themselves do not require labels, although score evaluation may still use training labels.

Framework	Pre-training	Multi-channel	Feature signal	Unlabeled structure	Label smoothness
h_{adj} + LI (Platonov et al., 2023a)	✓	✗	✗	✗	✓
CPM (Luan et al., 2024b)	✓	✗	joint	joint	joint
Tri-Hom (Zheng et al., 2024)	✓	✗	✓	partial	✓
NSA (ours)	✓	✓	✓	✓	✓

D ARCHITECTURE TRIAGE

The experiments above evaluate NSA primarily as a characterization framework. We now ask a secondary operational question: can dataset-level diagnostics help prioritize architectures before an expensive model search?

This question is deliberately separated from the controlled synthetic study in Appendix B. The earlier synthetic experiments validate whether NSA responds to known signal manipulations. Here, by contrast, the target is *architecture performance itself*. We therefore construct a separate cross-generator benchmark in which no architecture is designated as correct by construction and evaluate how well different dataset representations predict the empirical architecture ranking.

D.1 CROSS-GENERATOR ARCHITECTURE-SELECTION BENCHMARK

Each synthetic problem contains $n = 512$ nodes, three balanced classes, and eight-dimensional node features. We consider six architectures with relatively distinct signal-access mechanisms: MLP, GCN, GraphSAGE, APPNP, H2GCN, and GIN.

The benchmark contains six signal regimes: `feature_signal`, `smooth_diffusion`, `structured_heterophily`, `structural_role`, `neighbor_feature_context`, and `mixed_signal`. These regimes specify how predictive information enters the problem, not which architecture should win. The winner is determined empirically for every generated graph.

Training and test problems are generated by different graph-generation procedures while preserving the same underlying signal mechanism and broadly matched nuisance statistics. The resulting task

Table 16: Signal regimes and their train/test generator families. No architecture is assigned to a regime in advance; winners are always determined empirically.

Regime	Signal mechanism	Train gen.	Test gen.
Feature signal	Labels encoded in features; connectivity label-independent.	indep.-edge random	fixed-edge weighted
Smooth diffusion	Weak features, strong positive local smoothness on a sparse graph.	SBM	fixed-edge mixing
Structured heterophily	Heterophilic 1-hop neighborhoods; informative 2-hop structure.	latent-block SBM	fixed-edge latent-block
Structural role	Labels driven by class-dependent degree/role profiles; features uninformative.	expected-degree	fixed-edge role-weighted
Neighbor-feature context	Own features weak; neighborhood feature *composition* predictive.	node-wise preferential	fixed-edge compatibility
Mixed signal	Moderate feature and label–graph signal together.	SBM	fixed-edge mixing

is therefore a *cross-generator* architecture-selection problem rather than a random split of graphs produced by one generator.

The regimes are intentionally heterogeneous. Feature-signal problems place class information primarily in node attributes; smooth-diffusion problems contain weak direct feature signal but strong positive local label context; structured-heterophily problems contain misleading one-hop neighborhoods with informative higher-order structure; structural-role problems encode labels primarily through unlabeled graph roles; neighbor-feature-context problems require combining graph adjacency with the feature composition of a node's neighborhood; and mixed problems contain moderate signal through multiple channels. Table 16 summarizes each regime's signal mechanism together with its train and test generator families.

Table 17: Pre-specified architecture-only pilot criteria (36 held-out pilot problems).

Criterion	Required	Observed	Result
Largest architecture share	≤ 0.45	0.361	Pass
Distinct winners	≥ 4	6	Pass
Constant-selector regret	≥ 0.05	0.065	Pass
Median within-graph spread	≥ 0.08	0.279	Pass
Fraction with spread > 0.05	≥ 0.70	0.972	Pass
Fixed Top-3 hit rate	≤ 0.80	0.833	Fail
Median best–second gap	≥ 0.01	0.029	Pass

D.2 Pilot–Freeze–Final Protocol

Synthetic benchmark construction can easily introduce confirmation bias if the generators are repeatedly modified after observing the proposed diagnostic. We therefore separate architecture-level benchmark calibration from descriptor evaluation.

During an initial architecture-only pilot, generator parameters were assessed using only downstream architecture performance, winner diversity, model-selection headroom, and coarse graph statistics. NSA was inspected only for implementation sanity checks, such as numerical failures, saturation, or known descriptor leakage; NSA-versus-homophily selection performance was not used to tune the benchmark.

Before evaluating any descriptor-based selector, we fixed seven architecture-level acceptance criteria measuring winner concentration, selection headroom, architecture spread, and ambiguity, listed in Table 17. Six of seven criteria were satisfied on the 36-problem pilot. The sole miss was the fixed Top-3 hit rate (0.833 versus a target of 0.80). We retained the benchmark unchanged rather than modify the generators after observing this result.

The generator specification and architecture bank were then frozen. The final descriptor comparison uses fresh random seeds, with 30 training problems and 15 held-out test problems per regime, for 180 training and 90 test problems in total.

Table 18: Mean architecture accuracy on the training-generator corpus. "Spread" is the median best–worst gap within a problem; "Best–2nd" is the mean gap between the top two architectures.

Regime	MLP	GCN	SAGE	APPNP	H2GCN	GIN	Spread	Best–2nd
Feature signal	.731	.565	.738	.728	.734	.542	.225	.016
Smooth diffusion	.353	.600	.601	.713	.748	.741	.409	.028
Structured heterophily	.354	.460	.672	.421	.830	.796	.498	.041
Structural role	.356	.551	.508	.420	.354	.723	.402	.167
Neighbor-feature context	.352	.374	.455	.364	.439	.350	.145	.033
Mixed signal	.540	.679	.651	.655	.655	.700	.152	.018
Mean	0.448	0.538	0.604	0.550	0.627	0.642	–	–

D.3 Benchmark Validity

The architecture diversity observed during the pilot reproduces on the frozen final corpus. Mean oracle accuracy is 0.701, whereas the best constant architecture, GIN, achieves 0.630, leaving 7.18 percentage points of mean selection headroom.

No single architecture dominates the benchmark. GIN wins 42.2% of the 90 held-out problems, H2GCN 23.3%, GraphSAGE 21.1%, and MLP, GCN, and APPNP 4.4% each. Thus, all six architectures are optimal on at least some problems, and selecting the same model everywhere leaves substantial performance unrealized.

Architecture performance also varies substantially across regimes, as reported in Table 18. No single architecture is uniformly strong: feature-signal problems favor feature-centric or flexible models, structured heterophily favors architectures such as H2GCN and GIN, and neighbor-feature-context and mixed regimes produce more distributed performance profiles. These patterns emerge empirically from the benchmark rather than from a one-regime–one-architecture assignment.

The train and test generators are also broadly matched in coarse nuisance statistics, as shown in Table 19: average degree, degree variation, clustering, spectral gap, homophily, and edge-feature similarity remain in comparable ranges across generator families within each regime. This reduces the risk that cross-generator selection is driven primarily by generator identity rather than by the underlying signal mechanism.

Table 19: Train/test nuisance statistics (train / test mean).

Regime	Avg. deg.	Std. deg.	Edge hom.	Clustering	Spec. gap	Feat. cos.
Feature signal	14.4/12.5	3.7/3.9	.33/.33	.028/.024	.50/.46	.07/.07
Smooth diffusion	3.1/3.2	1.6/1.7	.80/.85	.008/.010	.00/.00	.00/.00
Structured het.	4.1/4.0	1.9/2.0	.04/.05	.002/.003	.02/.02	.00/.00
Structural role	8.5/8.9	5.5/6.0	.45/.44	.026/.032	.03/.05	−.00/−.00
Neighbor-feat. ctx.	8.0/7.8	2.0/2.9	.35/.34	.015/.017	.35/.33	.08/.04
Mixed signal	5.7/5.7	2.4/2.4	.56/.63	.012/.012	.17/.13	.05/.05

D.4 NSA and Homophily Are Complementary

We next predict the six-dimensional architecture-performance vector from dataset-level diagnostics. Every learned representation uses exactly the same pipeline: a StandardScaler fitted on the training-generator corpus followed by a multi-output Ridge regressor. The fitted selector is evaluated without adaptation on the held-out generator family.

We compare Node Homophily; Homophily-4, consisting of edge homophily, node homophily, adjusted homophily, and label informativeness; the three-dimensional NSA representation $(\text{FSS}, \text{TSS}, \text{LSS})$; and concatenations of NSA with either Node Homophily or Homophily-4. The standalone comparisons were pre-registered; the two concatenated representations were introduced after examining the standalone NSA and homophily results, and we therefore do not describe those two tests as pre-registered. The graph generators, architecture bank, architecture performances, and final train/test corpus had already been frozen, however, and were not changed for this analysis.

These representations describe different aspects of a graph-learning problem. Homophily statistics summarize label–adjacency relationships, whereas NSA separates feature accessibility, unlabeled structural information, and local label context. The combined representations therefore test whether these views contain complementary information rather than treating them as competing definitions of the same quantity.

For problem x, we define selection regret as

$$\mathcal{R}(x) = \max_{a \in \mathcal{A}} P_a(x) - P_{\widehat{a}(x)}(x),$$

where $P_a(x)$ is the held-out accuracy of architecture a and $\widehat{a}(x)$ is the architecture selected from the predicted performance vector. Regret is the primary metric because it measures the actual cost of an incorrect choice. Full results are reported in Table 4 in the main text.

NSA alone substantially improves over scalar Node Homophily, reducing mean regret from 6.89 to 2.63 percentage points. The richer Homophily-4 profile is a stronger standalone selector, with regret 1.62 points, demonstrating that conventional label–graph statistics already contain considerable architecture-selection information.

The most informative result is obtained by combining the two views. NSA + Node Homophily achieves the lowest mean regret, 1.18 points – a 5.71-point reduction from Node Homophily alone, the largest absolute improvement among the three comparisons – while NSA + Homophily-4 reaches 1.31 points and gives the strongest overall architecture ranking. The paired complementarity tests are summarized in Table 20. Adding NSA to Homophily-4 significantly reduces regret ($p = 0.0234$, paired one-sided Wilcoxon), the combined representation also improves over NSA alone ($p = 0.0106$), and adding NSA to scalar Node Homophily yields the largest reduction in regret of the three comparisons ($p < 10^{-4}$).

Table 20: Paired one-sided Wilcoxon tests (left has lower regret).

Comparison	p-value
NSA + Node Homophily $<$ Node Homophily	$< 10^{-4}$
NSA + Homophily-4 $<$ Homophily-4	0.0234
NSA + Homophily-4 $<$ NSA	0.0106

D.5 Where Does the Complementarity Come From?

Table 21 breaks down selection regret by regime, and clarifies why combining NSA with homophily is useful.

On `feature_signal`, Node Homophily is almost completely uninformative for model selection, producing 24.6 points of regret, whereas NSA reduces regret to 1.4 points. This is precisely the type of problem that FSS is designed to distinguish: architecture suitability is driven by feature accessibility rather than by label–adjacency alignment.

The reverse pattern appears on `neighbor_feature_context`. NSA alone incurs 9.0 points of regret, and this is also the regime with the largest cross-generator discrepancy among the six mechanisms. The regime is intentionally interaction-driven: neither a node's own features nor graph structure alone fully describes the relevant signal; what matters is the feature composition of its neighborhood. Adding a single Node Homophily coordinate to NSA reduces regret to approximately 0.1 points. The result provides a concrete example in which two individually incomplete summaries become highly informative when combined.

Because this regime is both the largest source of cross-generator discrepancy and the regime where standalone NSA is weakest, we report a post-hoc sensitivity analysis, not used for the primary result, excluding its 15 problems. As shown in Table 22, regret drops for every representation once this regime is removed, consistent with it being a genuinely harder regime rather than a source of bias for one representation over another. The key comparison is unaffected in the way that matters: NSA + Homophily-4 remains significantly better than Homophily-4 alone ($p = 0.0234$), while its advantage over NSA alone no longer holds ($p = 0.707$) – expected, since this is precisely the regime where the homophily coordinates compensate for NSA's marginal-channel limitation.

Table 21: Per-regime selection regret, in percentage points (pp). Lower is better.

Regime	NodeH	Hom-4	NSA	NSA+NodeH	NSA+Hom-4	constant selector
Feature signal	24.6	1.5	1.4	1.4	**1.2**	24.6
Smooth diffusion	4.9	2.0	**0.9**	**0.9**	**0.9**	**0.9**
Structured heterophily	0.7	**0.7**	1.0	1.0	1.1	4.8
Structural role	0.0	0.0	0.0	0.0	0.0	0.0
Neighbor-feature context	9.9	0.6	9.0	**0.1**	0.6	11.5
Mixed signal	**1.3**	5.0	3.5	3.7	4.0	**1.3**

On `structured_heterophily`, Homophily-4 performs particularly well, in part because label informativeness can capture structured dependence between endpoint labels even when same-label adjacency is rare.

The mixed regime is intentionally less clean: several architectures are competitive and the best constant selector is already strong. We therefore interpret the aggregate cross-regime result, rather than requiring NSA to be the strongest selector in every individual mechanism.

Table 22: Sensitivity to excluding `neighbor_feature_context` (post-hoc). Regret in percentage points (pp).

Representation	All 90	Excl. (75)
Node Homophily	6.9	6.3
Homophily-4	1.6	1.8
NSA	2.6	1.4
NSA + Node Homophily	1.2	1.4
NSA + Homophily-4	1.3	1.5
constant selector	7.2	6.3

Augmenting the representations with five generic graph statistics (mean and standard deviation of log-degree, clustering coefficient, normalized spectral gap, and mean edge-feature cosine similarity) does not improve on the homophily-based combinations above: Homophily-4 + Meta-5 reaches 1.30 points of regret and NSA + Meta-5 reaches 2.48 points, both comparable to or worse than the combinations already reported. We omit further detail since these generic features do not bear on the conceptual relationship between NSA and label–graph diagnostics.

D.6 Budgeted Architecture Search

Exact winner prediction is not always necessary in practice. A diagnostic can still be useful if it places high-performing architectures near the top of a shortlist. We therefore evaluate budgeted search by ranking the six architectures according to the predicted performance vector and measuring the best accuracy obtained after evaluating the top K recommendations.

The standalone representations already recover most of the oracle performance with very small budgets. With two architecture evaluations, NSA recovers 99.3% of oracle performance and Homophily-4 recovers 99.2%, compared with 98.0% for Node Homophily and 98.3% for the constant ranking. This complements the regret analysis: even when NSA does not identify the exact winner at $K = 1$, it often places a near-optimal architecture very high in its ranking Table 23.

Table 23: Budgeted search: % of oracle recovered after top-k recommendations.

Method	k=1	2	3	4	5	6
Homophily-4	97.7	99.2	99.4	99.7	99.9	100.0
NSA	96.3	99.3	99.7	99.8	100.0	100.0
Node Homophily	90.2	98.0	99.4	99.7	99.9	100.0
constant selector	89.8	98.3	99.5	99.7	99.9	100.0
Random	79.4	89.5	94.2	97.0	98.7	100.0

D.7 Generalist Architectures Reduce Selection Headroom

The specialist architecture bank above is designed to test whether differences in signal structure can support meaningful architecture triage. Modern flexible architectures pose a complementary question: what happens when a single model can exploit several signal types simultaneously?

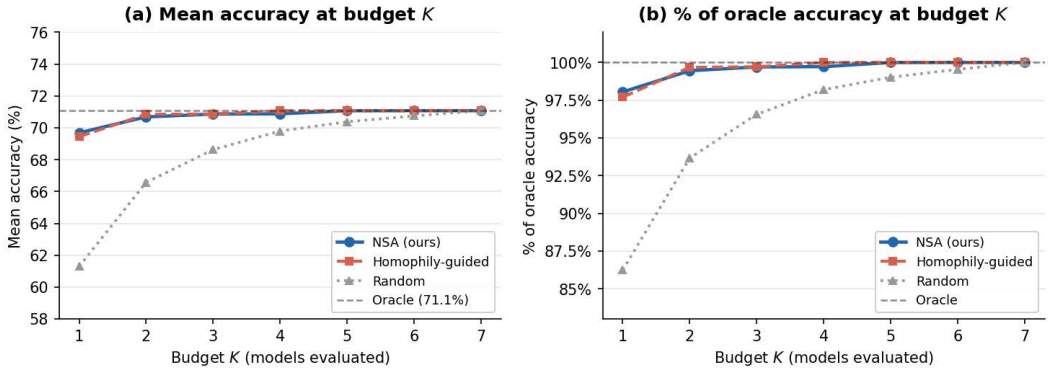

Figure 2: **Budgeted model search on real benchmarks (classical models only). Left:** mean test accuracy achieved after evaluating the first K models in each rule's ordering. **Right:** the same quantity normalized by oracle accuracy. NSA-guided search reaches near-oracle performance within very small budgets and substantially outperforms uninformed search in the low-budget regime.

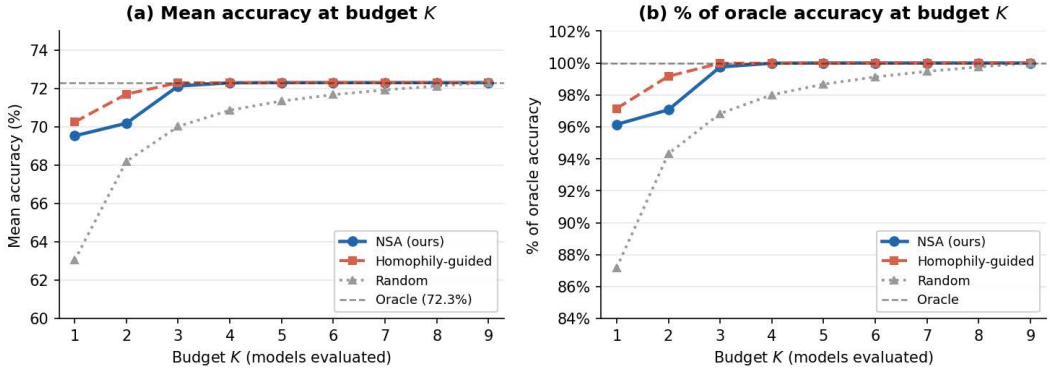

Figure 3: **Budgeted model search on real benchmarks (all nine models, including graph transformers).** This GT-inclusive setting is reported as a stress test. Because GT models occupy many of the top positions on this suite, the gap between different signal-aware rules is compressed, and even simple GT-first strategies can perform strongly.

Our real-benchmark experiments illustrate this stress-test regime. This analysis uses all real-world benchmarks excluding Genius and ogbn-arxiv, for which results are available for all nine models. As shown in Figure 3, when GPS and SGFormer are included alongside the classical models, *graph transformers occupy many of the top positions across the datasets*. Therefore, even a simple strategy that places the transformer family early can attain low regret, and the differences between dataset-aware guidance rules become compressed. This is not a failure of the diagnostics; it reflects reduced intrinsic selection headroom when one flexible family behaves as a strong generalist. The classical-model setting provides a cleaner view of signal-aware triage. As shown in Figure 2, restricting the search to the seven non-transformer models {MLP, LP, APPNP, GCN, GraphSAGE, H2GCN, SGC} makes the oracle distribution less concentrated. NSA- and homophily-guided search both substantially outperform random ordering at very small budgets and recover near-oracle performance after only one or two evaluations. The difference between the two signal-aware rules is small on these 20 real datasets, consistent with the limited size of the suite and the strength of homophily as a practical baseline.

We therefore view the two settings as answering different questions. The cross-generator specialist benchmark isolates whether dataset signal structure contains information relevant to model choice, while the GT-inclusive real benchmark tests how much practical selection value remains when strong general-purpose architectures are available.

Takeaway. The architecture-triage experiments do not suggest that NSA should replace validation or exhaustive model search when sufficient computation is available. Rather, they show that its three interpretable coordinates contain selection-relevant information beyond scalar node homophily and,

importantly, complement richer label–graph diagnostics. The practical value is greatest when only a small number of model evaluations are affordable and when the candidate architectures expose sufficiently different signal-access mechanisms.