
Graph Label Alignment: A Diagnostic Atlas for Graph Classification

Neelam Akula

Department of Mathematical Sciences
University of Texas at Dallas
Richardson, TX 75080
neelam.akula@utdallas.edu

Varun Shiralkar

Department of Computer Science
University of Texas at Dallas
Richardson, TX 75080
varun.shiralkar@utdallas.edu

Murat Kantarcioglu

Department of Computer Science
Virginia Tech
Blacksburg, VA 24061
muratk@vt.edu

Baris Coskunuzer

Department of Mathematical Sciences
University of Texas at Dallas
Richardson, TX 75080
coskunuz@utdallas.edu

Abstract

Graph classification benchmarks vary widely in difficulty: some are solved by simple pooled baselines, while others remain challenging for GNNs and graph transformers. We take a data-centric view and ask whether benchmark success is driven by genuine structure–feature reasoning or by label signal already accessible from restricted marginal views. We introduce *graph label alignment*, a lightweight diagnostic that measures label accessibility through two supervised probes computed without training a graph encoder: *structural label alignment* (SLA), based on permutation-invariant topological and spectral summaries, and *feature label alignment* (FLA), based on pooled node features with edges removed. Across real graph classification benchmarks and multiple pooled backbones, the resulting SLA–FLA plane strongly organizes test accuracy, with low-alignment datasets forming a consistent regime where standard pooled pipelines struggle. To probe what lies beyond marginal accessibility, we introduce controlled interaction-dominated stress tests, including heterophilic ego-graphs and synthetic *Twins* tasks whose structural and feature marginals are matched across classes but whose role–feature couplings differ. These results show that high alignment explains much of easy benchmark performance, while low alignment reveals regimes where success requires preserving structure–feature correspondence beyond standard pooled representations.

1 Introduction

Strong performance on a graph classification benchmark can mean two very different things: a model may exploit relational structure, or it may exploit label signal already accessible from coarse marginal views of the data. This ambiguity is increasingly important because graph classification benchmarks vary widely in difficulty: many are handled well by simple pooled baselines, while others challenge expressive GNNs and graph transformers [EPBM20, DJL⁺23]. This motivates a data-centric question: *Where does the label signal live in a graph classification dataset, and which parts of it are accessible to standard pooled graph models?*

We address this question through a supervised restricted-view diagnostic for graph classification. The goal is to separate marginal label accessibility from interaction-dependent signal: high marginal

accessibility can explain why many benchmarks are easy for standard pooled pipelines, while low marginal accessibility highlights regimes where success requires preserving structure–feature correspondence beyond coarse pooled summaries. If labels are predictable from structure alone or from the multiset of node features alone, then strong accuracy may largely reflect this accessibility rather than richer relational reasoning [LCS⁺24, CWSR25]. Conversely, when both restricted views are weak, remaining signal must come from higher-order structure–feature interactions, information not captured by the reference probes, or noise. We instantiate this low-marginal regime through controlled Twins tasks, where structural and feature shortcuts are eliminated by design.

Graph label alignment. Given a dataset $\mathcal{D} = \{(G_i, X_i, y_i)\}_{i=1}^N$, we introduce two probe-based scores that quantify marginal label accessibility without training a graph encoder. *Structural label alignment (SLA)* measures label predictability from a fixed, permutation-invariant structural descriptor of G_i alone, based on topological and spectral summaries. *Feature label alignment (FLA)* measures label predictability when edges are discarded and only the multiset of node features in X_i is retained. Together, $(\widehat{\text{SLA}}, \widehat{\text{FLA}})$ place datasets on an interpretable *graph label alignment plane*, enabling benchmark comparison and helping identify when performance is likely driven by marginal accessibility versus structure–feature interaction.

A plane that organizes pooled-model performance. Across real benchmarks and pooled graph classifiers, including GCN, GraphSAGE, GAT, and GPS, datasets with higher SLA or FLA tend to yield higher accuracy, while low-alignment datasets cluster in regimes where pooled backbones generally perform worse. Architecture choice shifts absolute performance, but does not reliably remove this difficulty. We quantify the trend using Pearson correlations between alignment scores and test accuracy, and use $\text{AlignAvg} = (\text{SLA} + \text{FLA})/2$ as a compact dataset ordering.

Controlled stress tests expose what the alignment plane misses. Real datasets confound domain, size, feature type, and class structure, so we complement them with controlled settings. A synthetic sweep independently dials structural and feature separation, validating that accuracy contours follow the alignment geometry. Heterophilic ego-graph classification provides realistic low-marginal tasks where pooled message-passing baselines achieve low accuracy, typically $\sim 25\text{--}45\%$, and performance is sensitive to hop radius. Most sharply, in balanced binary and three-class *Twins* tasks, classes have identical structural and feature marginals but distinct role–feature couplings; pooled GNNs and GPS remain near random guessing. These results suggest a correspondence bottleneck in standard pooled graph-classification pipelines: when label signal resides in global role–feature coupling, simply changing the encoder may not preserve the information needed for discrimination.

Contributions.

- We propose **graph label alignment**, a label-aware diagnostic for pooled graph classification, and define probe-based **SLA** and **FLA** scores for structure-only and feature-only restricted views.
- We show that pooled GNN and graph transformer performance on real benchmarks is strongly organized by the **SLA–FLA plane**, supporting AlignAvg as a compact dataset ordering.
- We provide **backbone-specific alignment predictors** and a practical **decision map** (Appendix G) that translate SLA/FLA regimes into backbone selection guidance.
- We introduce controlled **synthetic alignment sweeps** and **interaction-dominated stress tests**, including heterophilic ego-graphs and *Twins*, showing that low-marginal regimes can expose a failure mode of standard pooled pipelines under role–feature coupling.

2 Background

2.1 Related Work

Graph classification models and benchmarks. Graph classification is a core task in graph machine learning with applications in chemistry, biology, and social systems. Most methods learn node representations and aggregate them into graph-level embeddings, typically using message-passing GNNs with global pooling, or attention-based graph transformers with structural encodings. Evaluation is commonly performed on benchmark suites such as TU and OGB graph-level datasets

[MKB⁺20, HFR⁺20]. Prior work has emphasized that graph classification results can be highly sensitive to experimental protocols and that carefully tuned simple baselines can be competitive on several benchmarks [EPBM20].

Data-centric evaluation of graph classification datasets. A growing line of work asks whether popular benchmarks effectively test graph reasoning or whether they can be solved by simpler cues. [LCS⁺24] revisit graph classification benchmarks and propose effectiveness metrics based on performance gaps between graph-aware methods and simpler structure-only or attribute-only baselines, together with correlation-controllable synthetic datasets. [CWSR25] propose RINGS, a mode-perturbation framework that evaluates datasets through model sensitivity to structure and feature perturbations, using measures such as performance separability and mode complementarity. These works assess benchmark quality via performance gaps and perturbation sensitivity.

In contrast, our goal is not to curate benchmarks, but to provide a simple *label-centric* view of where predictive signal is accessible under restricted information views. We introduce two probe-based scores: *structural label alignment* (SLA), measuring label predictability from permutation-invariant structural summaries alone, and *feature label alignment* (FLA), measuring label predictability from node-feature pooling summaries when edges are discarded. Together, they define an interpretable alignment plane that separates structure-accessible, feature-accessible, and low-marginal-accessibility regimes without training full graph encoders.

Alignment between structure, features, and labels. Several works study alignment between graph structure and node features, typically for node-level learning, and use spectral criteria to guide rewiring or denoising. For example, [LSD25] propose jointly rewiring the graph and denoising features by aligning leading spectral spaces of graph and feature matrices. Our setting differs in that we study *graph-level* classification and define alignment directly in terms of *label predictability* under restricted views, rather than structure-feature agreement itself. We use the term *alignment* strictly in this predictive sense.

We further use synthetic and ego-graph datasets to study regimes where labels are accessible from structure, accessible from features, or suppressed under both marginal views. The low-marginal regime is especially useful diagnostically: it can indicate noise, probe underfitting, or genuinely interaction-dependent signal. Our controlled Twins construction isolates the last case by matching structural and feature marginals across classes while varying only role-feature coupling. This lets us compare datasets and architectures on a common plane while making explicit what the alignment scores do and do not establish.

2.2 Structural & Feature Probes

We quantify label accessibility under two restricted views: structure-only and feature-only. For each view, we construct lightweight, permutation-invariant summary representations that serve as standardized probes across datasets. These probes are intentionally simple and are used to define controlled diagnostics for SLA and FLA (Section 3); they are not designed to maximize predictive performance. Implementation details are provided in Appendix B.

Structural summaries. We represent each graph using a combination of topological, spectral, and statistical descriptors. For topology, we compute Betti-profile summaries derived from persistent homology over a filtration (e.g., node degree) [DW22, CA24] (Appendix F.1). For spectral structure, we include a small set of extremal eigenvalues of the normalized Laplacian and adjacency matrices. Together, these features provide compact, permutation-invariant summaries of graph structure (see Appendix B).

Feature-only summaries. For the feature-only view, we discard edges and treat node features as an unordered multiset. Each graph is represented by a permutation-invariant pooled descriptor; in our reference setting, we use the feature centroid. Alternative summaries are evaluated in Appendix B.

3 Graph–Label Alignment Framework

3.1 Problem Setup

We consider a graph classification dataset $\mathcal{D} = \{(G_i, X_i, y_i)\}_{i=1}^N$, where $G_i = (V_i, E_i)$ is a graph, $X_i \in \mathbb{R}^{|V_i| \times d}$ are node features, and $y_i \in \{1, \dots, C\}$ is a graph label. We view each sample through two restricted information views: **structure-only** (discard X_i) and **feature-only** (discard E_i), and ask how much label signal is accessible from each view via standardized, lightweight probe predictors. Throughout, we use *alignment* strictly to mean *label predictability under a restricted view*, rather than agreement between structure and features (see Section 2).

3.2 Restricted-View Summary Vectors

To quantify label signal under each view, we map each graph to a fixed-dimensional summary vector. These summaries are deliberately simple and permutation-invariant: they are not intended to maximize accuracy, but to provide controlled, comparable diagnostics across datasets.

Structural view. Node features are discarded and each graph G is represented by a *Structural Summary Vector* $\phi_{\text{struct}}(G) \in \mathbb{R}^{48}$ that concatenates three families of permutation-invariant descriptors: a degree-filtration Betti profile (PH, 20 dims), a compact spectral signature from the normalized Laplacian and adjacency matrix (8 dims), and global graph statistics including size, density, degree quantiles, and clustering moments (20 dims). The vector uses no attribute information; zero-padding is applied when fewer than four nonzero eigenvalues are available. Full construction details are in Appendix B; component-level ablations supporting their complementarity are in Appendix C.1.

Feature-only view. We discard edges and treat each graph as an unordered multiset of node features, and represent each graph by a *Feature Summary Vector* $\phi_{\text{feat}}(G) \in \mathbb{R}^d$ given by the feature centroid, $\phi_{\text{feat}}(F) = 1/|V| \sum_{v \in V} X_v$. Alternative feature summaries considered are reported in App. B.

3.3 Alignment Score Computation

Given summary vectors $\phi(G_i) \in \{\phi_{\text{struct}}(G_i), \phi_{\text{feat}}(G_i)\}$, we estimate label accessibility using a fixed probe protocol. We train two lightweight classifiers—logistic regression and an RBF-kernel SVM—using seeded, stratified cross-validation, and define each alignment score as the average of their cross-validated accuracies.

When $\phi = \phi_{\text{struct}}$, this defines the *Structural Label Alignment* score $\widehat{\text{SLA}}$, measuring how predictive graph structure alone is of the label. When $\phi = \phi_{\text{feat}}$, it defines the *Feature Label Alignment* score $\widehat{\text{FLA}}$, measuring how informative the marginal node-feature distribution is when adjacency is ignored.

Both scores are operational restricted-view diagnostics, not causally isolated measures of structural or feature contributions. If node features encode structural information, as atom types do in molecular graphs where valence partly determines local topology, FLA absorbs that signal by design. Thus FLA asks “how much label signal is accessible from features alone?”, not “how much signal is causally feature-driven?” Similarly, low SLA means structure is not predictive *under the chosen structural summary*, not that no structural signal exists.

Probe details are in Appendix B.3. To assess whether the reference protocol underestimates accessible signal on a given dataset, Appendix C.2 introduces a *probe sufficiency self-check* that compares reference scores against an alternative five-classifier suite and flags probe-sensitive datasets.

3.4 Using Alignment in Analysis

We use $(\widehat{\text{SLA}}, \widehat{\text{FLA}})$ as a two-dimensional summary of marginal label accessibility for pooled graph classification pipelines. In Section 6, we visualize datasets in the SLA–FLA plane and analyze pooled-model behavior via regime-based comparisons. We also use $\text{AlignAvg} = (\text{SLA} + \text{FLA})/2$ (when both are defined) as a one-number ordering of alignment for correlation analysis and plotting.

Two properties of this plane are worth noting upfront. First, SLA and FLA are structurally independent axes: the synthetic sweep in Section 4 confirms they decouple cleanly under controlled generation,

and several real datasets (e.g., MOLHIV, LETTER-HIGH, COIL-DEL) show meaningful SLA–FLA gaps. The positive correlation observed across the broader benchmark suite is an empirical property of existing collections, not a built-in property of the framework: in many current benchmarks, datasets with high marginal accessibility under one view also tend to exhibit high marginal accessibility under the other. Second, low alignment on both axes does not automatically imply interaction-driven signal: it means labels are not accessible from either marginal view, which can arise from genuine noise, from signal in higher-order structure–feature interactions, or from probe underfitting on atypical datasets (see Appendix C.2). The distinction matters for interpretation and is addressed directly in the controlled experiments of Section 5.

4 Synthetic Alignment Testbeds

We use controlled synthetic graph families for two purposes. First, an *alignment plane sweep* spans a range of SLA and FLA values and tests whether pooled-model performance follows the $(\widehat{\text{SLA}}, \widehat{\text{FLA}})$ geometry under controlled variation. Second, the *Twins* construction removes marginal shortcuts and isolates interaction-only signal: prediction is possible only if a model captures how node features are arranged relative to graph structure.

4.1 Alignment Plane Sweep

Goal. The sweep instantiates three regimes: structure-only (high SLA, low FLA), feature-only (high FLA, low SLA), and jointly aligned (high SLA, high FLA). By varying structural and feature signal independently, we obtain datasets whose location in the alignment plane can be controlled while other factors are fixed.

Base graph generator. For each graph with label $y \in \{1, \dots, K\}$, with $K = 3$, structure is generated from a label-controlled mixture. Let $s \in [0, 1]$ be the *structural separation* parameter. With probability s , the graph is sampled from a class-specific generator $\mathcal{G}_y$; with probability $1 - s$, it is sampled from a shared label-independent generator $\mathcal{G}_0$.

The class-specific generators are drawn from canonical random graph families such as Erdős–Rényi, Barabási–Albert, Watts–Strogatz, and two-block stochastic block models, with class-dependent parameters chosen to induce structural differences. The shared generator $\mathcal{G}_0$ is an Erdős–Rényi model with fixed edge probability. We add label-agnostic edge rewiring noise by rewiring a fixed fraction p_{rewire} of edges uniformly at random, and sample graph sizes from a fixed distribution so expected node counts are matched across classes. Thus $s \approx 0$ produces structure-agnostic datasets, while $s \approx 1$ produces structure-informative datasets. Full parameter values are in Appendix E.

Feature generator. Node features are generated independently of graph size. For each node v , we sample $x_v \sim \mathcal{N}(0, \sigma^2 I)$ with fixed σ^2 . Let $f \in [0, 1]$ be the *feature separation* parameter. With probability f , all nodes in a graph of label y receive a class-dependent mean shift μ_y ; otherwise features remain label-independent noise. The class means are sampled once, normalized, and scaled by a fixed magnitude. Thus increasing f makes pooled feature summaries more label-predictive, while $f \approx 0$ matches marginal feature distributions across classes.

Regime instantiations. We sweep (s, f) over a fixed grid to cover the SLA–FLA plane. This yields three families: *Structure-only* (vary s with $f = 0$), *Feature-only* (vary f with $s = 0$), and *Jointly aligned* (set $s = f$). Grid size and graphs per cell are given in Appendix E.

4.2 Twins Construction

Goal. Twins eliminate structure-only and feature-only shortcuts. By construction, all classes share the same graph distribution and the same multiset of node feature types, so SLA and FLA are near chance. Label signal comes only from how feature types are coupled to structural roles.

Per-graph labeled variants. For each base graph $G = (V, E)$ sampled from a fixed-parameter SBM [HLL83], we generate multiple labeled variants that share the same edge set E and the same balanced multiset of feature types. Each variant contains exactly $n/3$ nodes of each prototype type,

so the marginal feature distribution is matched across labels. Variants differ only in the assignment of feature prototypes to nodes, forcing successful models to detect role–feature coupling rather than structure-only or feature-only signal.

Role scores and quantile bins. To induce structured feature placement, we compute a scalar role score $s(v)$ for each node and partition nodes into three equal-sized bins by tertiles. We consider four scoring functions, detailed in Appendix E: **Linear**, **Radial**, **PageRank**, and **Eigen**. Feature prototypes (μ_1, μ_2, μ_3) are then assigned to the bins in class-specific orders with additive Gaussian noise, while preserving identical prototype counts across classes.

Classes. Label 0 corresponds to a random assignment obtained by uniformly permuting prototype labels across nodes while preserving the $n/3$ - $n/3$ - $n/3$ counts. Labels 1, 2, ... correspond to structured assignments that align prototype types with role-score bins. We consider binary Twins (random vs. one structured assignment) and three-class Twins (random vs. two structured assignments). Since both structure and feature marginals are matched, any gain over chance must come from preserving global role-feature correspondence.

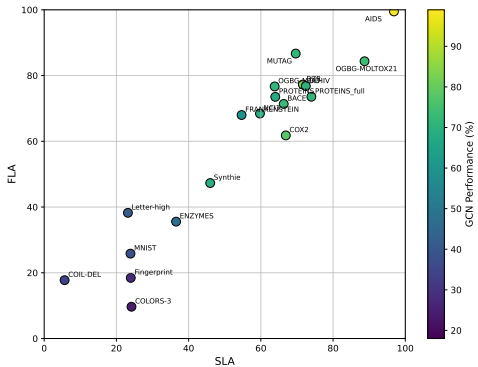
5 Experiments

Overview and evaluation protocol. Our goal is to characterize how *pooled* graph classifiers behave as a function of graph-label alignment, and to separate two questions: (i) whether marginal label accessibility (SLA/FLA) organizes performance on *standard* graph benchmarks, and (ii) what fails when labels depend primarily on *structure-feature interaction* that permutation-invariant readouts may wash out. We compute SLA/FLA *once per dataset* using the restricted-view probes from Section 3; these scores are model-agnostic and require no GNN training. To answer (i), we use **general benchmarks** spanning diverse domains and naturally covering a wide range of SLA/FLA: **(1a)** real graph classification datasets, and **(1b)** a synthetic alignment-plane sweep, where we can dial structural and feature signal continuously while holding other factors fixed. To answer (ii), we use **interaction-driven stress tests** designed to suppress marginal shortcuts and stress coupling: **(2a)** heterophilic node-to-graph ego networks (Appendix D), which provide realistic interaction-dominated instances, and **(2b)** synthetic *Twins*, which match structure and feature marginals across classes by construction and isolate role-feature coupling as the only discriminative signal.

For real benchmarks and ego-graphs, we follow the standard dataset splits and evaluation protocols; for synthetic datasets, we use balanced classes with a 60/20/20 train/validation/test split (Section A). Our code is available at the link: <https://github.com/avp-neelam/GraphLabelAlignment>

Real benchmark suite (1a). We use a diverse set of graph classification datasets spanning molecules, proteins, social networks, and image-to-graph constructions (TU-style and OGB molecular graphs). Some datasets have node features (SLA and FLA); others are featureless (FLA undefined), in which case we report only SLA and omit AlignAvg. Dataset lists, statistics, and feature availability are in Appendix Table 4.

Synthetic alignment sweep (1b). To control alignment, we generate synthetic datasets by sweeping structural and feature separation (s, f) as in Section 4, covering structure-only, feature-only, and jointly-aligned regimes. We evaluate the same architectures and tuning protocol as on real benchmarks and compute SLA/FLA using the restricted-view probes.



Heterophilic ego-networks (2a). To model interaction-driven classification, we convert heterophilic node benchmarks into ego-graph tasks: for each node, we extract the induced k -hop ego-graph and use

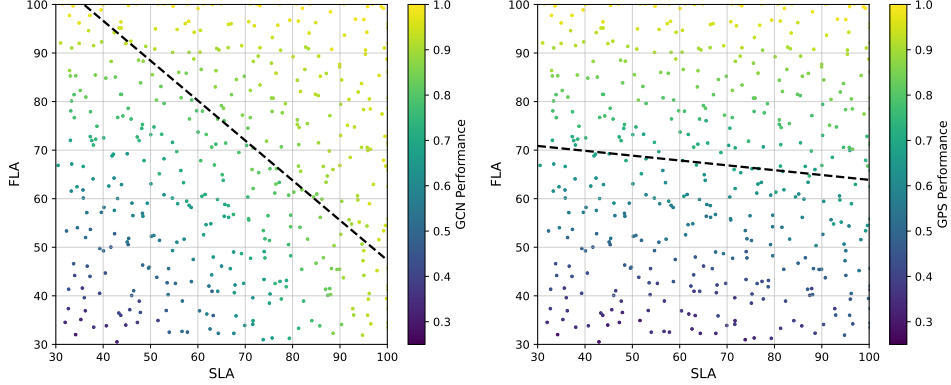


Figure 2: **Synthetic alignment sweep (GCN& GPS)**. Each point is a synthetic dataset positioned by its SLA and FLA, colored by GCN(left) and GPS (right) accuracy. The dashed line marks an illustrative 80% accuracy contour. Other backbones are shown in Appendix Figure 5.

the center-node label as the graph label. We include a center-node indicator feature in all ego-graph experiments. Construction and preprocessing are in Appendix Section D.

Twin synthetic datasets (2b). We use Twins (Section 4) to remove structure-only and feature-only shortcuts while preserving the interaction-only label signal. All classes share the same SBM distribution and feature multiset; labels differ only by role-based feature assignment (Linear, Radial, PageRank, Eigen) versus random permutation. We evaluate binary and three-class settings. Unless noted, each dataset has 1000 graphs with 100 nodes; details are in Appendix E.

Models and baselines. We evaluate pooled graph classifiers: **GCN**, **GraphSAGE (SAGE)**, **GAT**, and **GPS**. Each model produces node embeddings and applies a permutation-invariant readout, mean pooling by default, to obtain graph-level embeddings for classification; readout variants and hyperparameters are in Appendix Section A. For Twins, we also evaluate **GCN+DiffPool** to test whether hierarchical pooling mitigates the interaction-only failure mode.

6 Results

6.1 Alignment plane on real benchmarks

We first test whether our restricted-view alignment scores organize the performance of standard *pooled* graph classifiers on real graph classification benchmarks. For each dataset, we report baseline accuracies for **GCN**, **SAGE**, **GAT**, and **GPS** together with the corresponding alignment scores (**SLA** and **FLA** when node features are available; **SLA** only for featureless datasets). (Implementation details for backbones, pooled readout, and the hyperparameter grid are in Appendix Section A. Full dataset statistics are in Appendix Table 4.)

Key observation: a clear trend in the SLA–FLA plane. Across featureful benchmarks, datasets with higher SLA and/or higher FLA tend to achieve higher pooled-model accuracy, while low-alignment datasets cluster in a region where all pooled architectures struggle. This trend is visually apparent in Figure 1, which places each dataset in the SLA–FLA plane and colors points by **GCN** accuracy. Importantly, the qualitative organization is not driven by a particular backbone: corresponding plots for **SAGE**, **GAT**, and **GPS** show the same large-scale pattern (Appendix Figure 3). Throughout, we treat this as a descriptive organization of benchmark difficulty, rather than a causal claim about model behavior.

Table 1: Pearson correlations between alignment scores and model accuracy on 19 real benchmarks (datasets where FLA and AlignAvg ((SLA+FLA)/2) are defined; see Table 2). Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

ALIGNMENT	GCN	SAGE	GAT	GPS
SLA	0.935***	0.802***	0.941***	0.835***
FLA	0.928***	0.715***	0.961***	0.759***
ALIGNAVG	0.944***	0.769***	0.962***	0.808***

Table 2: **Real Graph Benchmarks.** Pooled-model accuracy (%) tracked against the alignment plane: structural and feature alignment scores (SLA/FLA, %). AlignAvg summarizes alignment as (SLA + FLA)/2. Datasets are ordered by decreasing AlignAvg when FLA is defined, and by decreasing SLA for featureless datasets (FLA/AlignAvg shown as –).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG↓
AIDS	98.40±0.30	98.30±0.51	95.95±1.04	98.85±0.30	99.51	96.85	98.18
MOLTOX21	73.45±0.41	73.63±0.53	71.95±0.40	73.70±1.01	84.35	88.71	86.53
MUTAG	71.28±5.48	69.74±4.10	67.18±5.48	70.77±7.36	86.69	69.65	78.17
DD	70.30±2.06	70.55±3.08	69.03±3.36	71.90±2.65	76.81	72.45	74.63
BZR	81.48±3.58	80.00±3.61	83.46±4.66	83.21±4.39	77.28	71.63	74.46
PROTEINS_FULL	71.07±3.16	71.70±3.73	73.84±1.64	74.11±2.06	73.56	73.99	73.78
MOLHIV	71.41±1.49	74.47±1.95	73.39±1.24	71.26±1.60	76.69	63.82	70.26
BACE	72.76±1.87	77.17±0.53	60.53±2.25	69.74±2.21	71.37	66.33	68.85
PROTEINS	67.59±3.50	68.12±2.22	67.05±4.72	68.93±3.62	73.52	63.99	68.76
COX2	78.09±5.37	78.30±4.44	76.81±5.23	77.87±5.27	61.79	66.91	64.35
NCII	68.30±1.59	69.39±1.20	64.53±1.81	69.39±0.88	68.45	59.75	64.10
FRANKENSTEIN	59.12±1.29	62.07±1.52	61.57±1.68	60.53±1.47	67.99	54.66	61.33
SYNTHE	69.00±5.83	77.00±4.00	45.50±4.91	45.00±4.47	47.30	45.98	46.64
ENZYMES	47.50±3.03	50.33±5.31	43.00±4.49	50.67±1.43	35.56	36.53	36.05
LETTER-HIGH	41.38±1.70	79.29±1.80	30.80±2.47	81.51±1.36	38.23	23.19	30.71
MNIST	37.85±0.70	48.80±0.61	32.01±0.76	47.33±0.62	25.83	23.90	24.87
CIFAR-10	27.77±0.59	39.85±1.06	31.70±0.39	41.54±0.72	18.45	23.98	21.22
COLORS-3	24.92±1.34	26.08±1.09	25.08±0.78	25.03±0.78	9.67	24.18	16.93
COIL-DEL	33.97±0.92	55.59±0.59	18.85±1.42	46.08±2.96	17.75	5.66	11.71
REDDIT-B	72.25±2.98	50.15±1.14	50.65±0.94	50.15±1.14	89.01	–	–
COLLAB	81.00±0.06	79.10±0.05	69.30±0.03	51.96±1.11	73.61	–	–
IMDB-B	63.30±0.04	58.90±0.06	54.60±0.07	50.00±2.30	70.43	–	–
REDDIT-5k	42.50±0.02	46.60±0.02	48.70±0.01	18.90±1.30	55.37	–	–
IMDB-M	40.20±0.05	42.00±0.05	37.70±0.05	31.20±3.77	46.00	–	–

Quantifying the trend via correlations. To quantify how well alignment predicts pooled accuracy beyond the qualitative plane plots, we compute Pearson correlations between each alignment score and model accuracy on the subset of real benchmarks where **FLA** (and thus **AlignAvg**) is defined.

All four backbones show strong positive correlations with both **SLA** and **FLA**: labels that are more accessible from either restricted view tend to yield better pooled-model performance. Moreover, $\text{AlignAvg} = (\text{SLA} + \text{FLA})/2$ is consistently among the most correlated summaries, providing a compact ordering for reporting and visualization. Since alignment scores are computed without training any GNN, these correlations serve as an architecture-agnostic diagnostic, suggesting that marginal label accessibility explains a substantial part of benchmark difficulty for these pooled models. Exact summary constructions, probe protocols, and complexity details are given in Appendix B.

Robustness of the alignment plane . Appendix C tests the design choices behind the scores. Removing structural blocks keeps SLA rank correlation with the full descriptor at 0.968 or above, and a 12-dimensional subset without persistent homology reaches 0.988. Richer feature summaries raise FLA on some low-alignment datasets, but keep rank correlation with centroid FLA at 0.934 or above. Across summaries and probes, no dataset moves directly between the high/high and low/low regimes, and correlations with accuracy keep their pattern across all four backbones. A stronger five-model probe suite flags LETTER-HIGH as the main probe-sensitive case. Appendix H reports compute cost, leave-one-dataset-out prediction of GCN accuracy (Pearson $r = 0.917$), a benchmark triage score, and scaffold splits, under which our correlations should be read as an in-distribution diagnostic.

6.2 Synthetic sweep validates the alignment plane

Real datasets vary simultaneously in many confounding factors (domain, size, feature type, and class structure), so we next test the alignment story in a controlled environment. We generate a grid of synthetic datasets by sweeping structural and feature separation parameters as described in Section 4, thereby spanning structure-only, feature-only, and jointly-aligned regimes while holding other factors fixed. (Full generation details and how SLA/FLA track the control parameters are provided in Appendix E, with full sweep tables reported at the end of the appendix.)

Table 3: **Twin setting** Accuracy results (%) for binary (random assignment vs. one structured assignment) and triclass tasks (random assignment plus two structured assignments).

	BINARY				TRICLASS					
	LINEAR	RADIAL	PAGERANK	EIGEN	LIN & RAD	LIN & PR	LIN & EIGEN	RAD & PR	RAD & EIGEN	PR & EIGEN
FLA	24.73	24.73	24.73	24.73	12.02	12.02	12.02	12.02	12.02	12.02
SLA	27.27	27.27	27.27	27.27	14.11	14.11	14.11	14.11	14.11	14.11
GCN	46.92	46.92	46.17	49.83	30.67	30.44	31.06	32.44	31.33	32.00
SAGE	49.50	48.50	48.67	49.50	33.11	32.39	31.56	32.00	33.00	33.11
GAT	47.92	45.58	44.00	46.83	31.78	28.56	31.00	31.06	31.39	31.33
GPS	48.50	49.67	49.25	49.33	32.06	32.50	32.94	32.67	32.83	32.22
DIFFPOOL	43.00	45.25	45.75	46.25	27.78	30.22	27.56	29.44	28.72	28.94

Key observation: accuracy contours follow the alignment geometry. Figure 2 visualizes the synthetic sweep in the alignment plane for GCN. The dashed curve marks an *illustrative* boundary for “good” performance (80% accuracy): datasets above this contour concentrate in high-alignment regions of the plane, while low-alignment regions remain challenging. As in the real benchmarks, the same qualitative organization is observed for other backbones (Appendix Figure 5), indicating that the alignment plane captures a stable performance geometry rather than an artifact of a single model.

6.3 Twin datasets expose a pooling limitation

The preceding results show that pooled performance is largely organized by *marginal* label accessibility. We now ask whether low alignment can also reflect a genuine interaction-dependent signal rather than noise. Twins answer this in a controlled setting: by construction, marginal structural and feature shortcuts are removed, while role–feature coupling remains discriminative.

Synthetic Twins isolate an interaction-only failure mode. We study *Twins* (Section 4), where all classes share the same SBM structure distribution and the same multiset of node feature types. Balanced binary and three-class variants differ only in how feature types are assigned to nodes according to a role score (Linear, Radial, PageRank, Eigen) versus a random permutation. Thus structure-only and feature-only probes are non-informative by design, and any gain over chance requires preserving global role–feature correspondence. Construction details, role-score definitions, and alignment suppression checks are given in Appendix Sections E.5 to E.8.

The result is clear: across all role-score variants and both binary and three-class settings, **pooled GNNs and GPS remain near random guessing** (50% binary, 33% three-class), with the failure most pronounced in the three-class setting. To test whether hierarchical pooling mitigates this effect, we also evaluate **GCN+DiffPool**; it remains near chance across Twins variants. These results suggest a correspondence bottleneck in standard pooled graph-classification pipelines: progress in this regime likely requires readouts or objectives that preserve the relation between node roles and feature assignments.

Heterophilic ego-graphs as realistic interaction-driven instances. We convert heterophilic node-classification graphs into ego-graph classification tasks (Appendix D) as a noisier but more realistic corroboration. These datasets are *interaction-dominated*: the class signal can depend on the arrangement of features around the center node rather than on global structure or pooled feature composition alone. Results are summarized in Table 15 and analyzed in Appendix D.1. Pooled message-passing baselines achieve low accuracy throughout, typically $\sim 25\text{--}45\%$, consistent with the Twins finding. GPS is a consistent outlier on WebKB at small k , achieving up to 60% on Cornell at $k = 1$ despite $\text{AlignAvg} \approx 25$. We discuss this boundary case in Appendix D.1, where global attention and center-node information may capture signal not reflected by marginal alignment scores.

Together, Twins and ego-graphs show that low alignment need not mean absence of signal: it can mark regimes where the discriminative information lies in structure–feature correspondence that standard pooled pipelines do not reliably preserve.

6.4 Discussion and takeaways

Overall, our experiments support the following takeaways.

Alignment is a practical organizer and diagnostic of pooled-model learnability. On real benchmarks, pooled-model accuracy is strongly structured by the SLA–FLA plane, and the same geometry appears under controlled synthetic sweeps where structural and feature signal can be dialed independently. This makes alignment a simple, architecture-agnostic way to diagnose where label signal is accessible from restricted views.

Low alignment is not fixed by swapping encoders. When both marginal views are weak, all pooled architectures degrade substantially and can approach near-chance behavior, suggesting that encoder choice alone is insufficient in interaction-dominated regimes; the bottleneck can lie in what standard pooled pipelines preserve.

Interaction-only tasks reveal a correspondence bottleneck. Heterophilic ego-graphs and Twins instantiate regimes where labels depend on how features are arranged relative to structure. In Twins, where marginal shortcuts are eliminated by design, pooled models remain near chance, indicating that permutation-invariant readout can discard global coupling signatures. This motivates alignment-aware evaluation and, in interaction-dominated settings, objectives or readouts that better preserve role-level structure–feature coupling beyond standard pooling. (For an optional backbone-aware alignment predictor and a practical decision map tying SLA/FLA regimes to backbone choice, see Appendix G.)

7 Conclusion

This work takes a data-centric view of graph classification by studying where label signal is accessible before training a graph encoder. Through *graph label alignment*, we show a consistent pattern across real benchmarks, heterophilic ego-graphs, and controlled synthetic families: when labels are accessible from structure or pooled features, standard pooled pipelines tend to perform well, whereas low alignment marks regimes where changing the encoder alone yields limited gains. Our interaction-focused stress tests further reveal a correspondence bottleneck in standard pooled pipelines, where global feature-role coupling can be lost despite expressive message passing. More broadly, these results suggest that graph benchmark analysis should distinguish marginal shortcuts from interaction-dependent signal. Future directions include readouts and objectives that preserve structure–feature correspondence, alignment-aware benchmark design, and extensions of this diagnostic view to link prediction, retrieval, and transfer.

Acknowledgments and Disclosure of Funding

This work was partially supported by National Science Foundation under grant DMS-2229417, OAC-2115094, CNS-2331424, ITE-2452833, Simons Foundation under grant SFM-21269, ARL/Army Research Office awards W911NF-24-1-0202 and W911NF-24-2-0114, and Virginia Commonwealth Cyber Initiative grants. The authors acknowledge the Texas Advanced Computing Center (TACC) at UT Austin for providing computational resources that have contributed to the research results reported within this paper.

References

- [A⁺23] Dashti Ali et al. A survey of vectorization methods in topological data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [BSFF⁺25] Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, et al. Position: Graph learning will lose relevance due to poor benchmarks. In *International Conference on Machine Learning*, 2025.
- [CA24] Baris Coskunuzer and Cüneyt Gürçan Akçora. Topological methods in machine learning: A tutorial for practitioners. *arXiv preprint arXiv:2409.02901*, 2024.
- [Chu97] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- [CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

- [CWSR25] Corinna Coupette, Jeremy Wayland, Emily Simons, and Bastian Rieck. No metric to rule them all: Toward principled evaluations of graph-learning datasets. In *ICML*, 2025.
- [DJL⁺23] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48, 2023.
- [DW22] Tamal Krishna Dey and Yusu Wang. *Computational Topology for Data Analysis*. Cambridge University Press, 2022.
- [EPBM20] Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *ICLR*, 2020.
- [HFR⁺20] Weihua Hu, Matthias Fey, Hongyu Ren, Masatoshi Nakata, Yuxiao Dong, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 2020.
- [HLL83] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic block-models: First steps. *Social networks*, 5(2):109–137, 1983.
- [HYL17] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *NeurIPS*, 30, 2017.
- [KW17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ICLR*, 2017.
- [LCS⁺24] Zhengdao Li, Yong Cao, Kefan Shuai, Yiming Miao, and Kai Hwang. Rethinking the effectiveness of graph classification datasets in benchmarks for assessing GNNs. In *IJCAI*, 2024.
- [LSD25] Jonas Linkerhäger, Cheng Shi, and Ivan Dokmanić. Joint graph rewiring and feature denoising via spectral resonance. In *ICLR*, 2025.
- [MKB⁺20] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*, 2020.
- [PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.
- [R⁺22] Ladislav Rampášek et al. Recipe for a general, powerful, scalable graph transformer. In *NeurIPS*, 2022.
- [Sei83] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3):269–287, 1983.
- [Spi11] Daniel A. Spielman. Spectral graph theory and its applications. <http://cs-www.cs.yale.edu/homes/spielman/spectral/>, 2011. Tutorial presented at IEEE FOCS 2011.
- [VCC⁺18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

Appendix

Appendix roadmap.

Section A gives **reproducibility** details, plus dataset metadata and per-backbone SLA–FLA plots.

Section B defines **SLA/FLA** summaries, probes, and complexity.

Section C reports **structural and feature ablations**, a **probe sufficiency self-check**, and **regime and seed stability**.

Section D covers **ego-graphs** (construction and extra results).

Section E details **synthetics** (sweep, coverage, Twins).

Section F reviews **signatures** (topology and spectrum).

Section G reports **GLA** (predictor, decision map, backbone rules), with full sweep tables at the end.

Section H reports **screening use** (compute cost, out-of-sample prediction, Δ_{graph}) and **scaffold splits**.

A Further Experimental Details

Implementation Details. We evaluated performance across two families of GNNs: (i) message-passing GNNs (GCN, GraphSAGE, and GAT); and (ii) a hybrid local-global GPS architecture. All models operate on node features, producing predictions via global pooling.

Message-passing GNNs. GCN [KW17], GraphSAGE [HYL17], and GAT [VCC⁺18] use a unified l -layer message-passing architecture with hidden dimension h . GCN and SAGE layers map using mean aggregation. For GAT, hidden layers use multi-head attention with H heads, yielding intermediate representation with dimension h . Each convolution is followed by batch normalization, a nonlinearity (ReLU for GCN/SAGE, ELU for GAT), and dropout with rate p . Node embeddings are aggregated using global mean pooling and passed through a linear classifier after the final layer.

GPS. GPS architecture [R⁺22] combines local message passing and global attention within each layer. Input features are first projected to dimension h via a linear map. Each of the l layers consist of a GraphSAGE convolution as the local MPNN component together with multi-head global attention with H heads. Outputs are normalized, passed through ReLU nonlinearity, and dropout with rate p . Residual connections are applied at each layer, updating node states. The final embeddings are pooled at the graph level and mapped to class logits using a linear classifier.

Grid Search for Hyperparameters. For real and heterophilic ego-graph datasets we run a grid search over hyperparameters: $l = 2, 3$ layers; $h = 32, 64, 128$ hidden dimension; $H = 1, 2, 4, 8$ multi-head attention (when relevant); $p = 0.3, 0.5$ dropout; and learning rate of 1×10^{-3} with early stopping on validation

B Structural and Feature Label Alignment

This appendix specifies the exact construction and implementation details of the *Structural Summary Vectors* and *Feature Summary Vectors* used to compute **Structure Label Alignment (SLA)** and **Feature Label Alignment (FLA)**, along with their computational complexity.

Let $\mathcal{D} = \{(G_i, X_i, y_i)\}_{i=1}^N$ denote a graph classification dataset, where each graph $G_i = (V_i, E_i)$ has node features $X_i \in \mathbb{R}^{|V_i| \times d}$ and class label y_i .

B.1 Structural Summary Vectors

For each graph $G = (V, E)$, we construct a fixed-dimensional *Structural Summary Vector* $\phi_{\text{struct}}(G) \in \mathbb{R}^{48}$ by concatenating topological, spectral, and global structural statistics. Formally:

$$\phi_{\text{struct}}(G) = [\beta_{0,1}(G) \parallel \lambda(G) \parallel \mu(G) \parallel g(G)]$$

Table 4: Dataset statistics across molecular, computer vision, synthetic, and social graph benchmarks.

GROUP	DATASET	#GRAPHS	AV.NODES	AV.EDGES	# FEAT.	#CLASSES
MOLECULAR	AIDS	2000	15.69	32.39	42	2
	BACE	1513	34.09	73.72	9	2
	BZR	405	35.75	76.72	56	2
	COX2	467	41.22	86.89	38	2
	DD	1178	284.32	1431.32	89	2
	ENZYMES	600	32.60	124.27	21	6
	FRANKENSTEIN	4337	16.90	35.75	780	2
	MOLHIV	41127	25.51	54.94	9	2
	MOLTOX21	7831	18.57	38.59	9	2
	MUTAG	188	17.93	39.59	7	2
	NCI1	4110	29.87	64.60	37	2
	PROTEINS	1113	39.06	145.63	4	2
	PROTEINS_FULL	1113	39.06	145.63	32	2
VISION	CIFAR-10	45000	117.65	941.18	3	10
	COIL-DEL	3900	21.54	108.47	2	100
	LETTER-HIGH	2250	4.67	9.00	2	15
	MNIST	55000	70.56	564.50	1	10
SYNTHETIC	COLORS-3	10500	61.31	182.05	5	11
	SYNTHIE	400	95.00	345.86	15	4
SOCIAL	COLLAB	5000	74.49	4914.43	0	3
	IMDB-B	1000	19.77	193.06	0	2
	IMDB-M	1500	13.00	131.87	0	3
	REDDIT-BINARY	2000	429.63	995.51	0	2
	REDDIT-5K	4999	508.52	1189.75	0	5

where $\beta_{0,1}(G)$ acts as the topological summary; $\lambda(G)$ and $\mu(G)$ the spectral summary; and $g(G)$ contains global structural statistics.

Topological Summary. For each graph we utilize sublevel-filtration based PH with degree as the filtration function, on ten quantiled bins, and compute the Betti-0 and Betti-1 numbers. To improve computational complexity we compute β_0 as the number of connected components and β_1 by Euler characteristic. The resulting concatenation

$$\beta_{0,1}(G) = (\beta_0(G^1), \beta_0(G^2), \dots, \beta_0(G^{10}), \beta_1(G^1), \beta_1(G^2), \dots, \beta_1(G^{10}),) \in \mathbb{R}^{20}$$

is the defined topological summary included in $\phi_{\text{struct}}(G)$.

Spectral Summary. Let A be the adjacency matrix of G and $L = I - D^{-1/2}AD^{-1/2}$ the normalized Laplacian. To compute $\lambda(G) \in \mathbb{R}^4$ we find the four smallest (nonzero) eigenvalues of L , similarly $\mu(G) \in \mathbb{R}^4$ are the four largest eigenvalues of A . To maintain fixed length summary vectors across graphs, if a graph has fewer than four eigenvalues all available values are used and the remaining length is padded with zeros [Chu97].

Global Summary. The global summary vector $g(G) \in \mathbb{R}^{20}$ contains number of nodes, number of edges, graph density, and degree quantiles at 5%, 25%, 50%, 75%, and 95%. Additionally, we report mean, standard deviation, skewness, and excess kurtosis for node-level degree, clustering coefficient, and triangle count.

B.2 Feature Summary Vectors

For each graph G with present node feature matrix $X \in \mathbb{R}^{|V| \times d}$, we construct a *Feature Summary Vector* $\phi_{\text{feat}}(G) \in \mathbb{R}^d$. Our reference summary is the centroid of the corresponding feature-space point-cloud. That is,

$$\phi_{\text{feat}}(G) = \frac{1}{|V|} \sum_{v \in V} X_v.$$

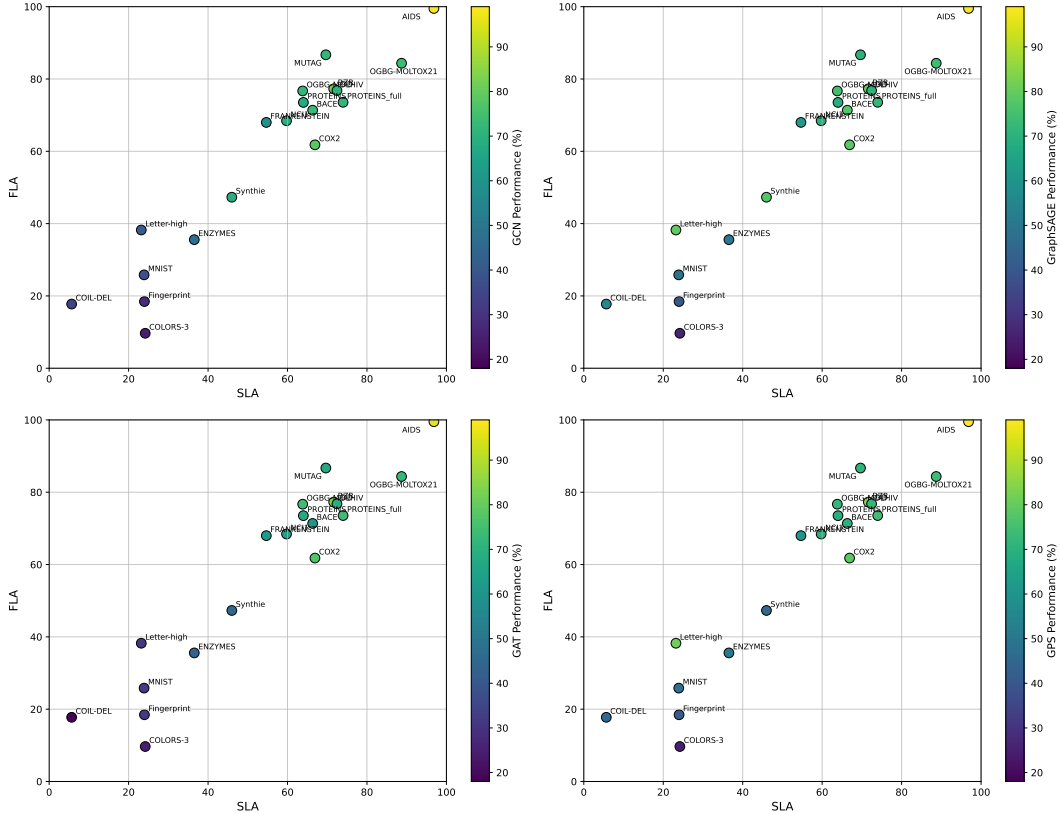


Figure 3: **GNN on graph classification benchmarks.** Each point is a real benchmark dataset in the SLA-FLA plane, colored by GNN classification accuracy. Beginning in the top left: GCN, GraphSAGE, GAT, and GPS.

We use the centroid because it is the cheapest summary and its FLA relates to GNN accuracy about as strongly as the richer summaries do, not because it scores highest. Richer permutation-invariant summaries often raise FLA itself; Section C.3 reports this comparison.

B.3 Alignment Score Computation and Model Choice

We detail the construction and evaluation protocol for the structural and feature alignment scores introduced in Section 3. From the fixed-length summary vectors ϕ_{struct} and ϕ_{feat} , we obtain two dataset-level design matrices X_S and X_F , with class labels y inherited from the original graph classification task.

Baseline probe suite. In the main experiments, we consider a fixed suite of two elementary classifiers: logistic regression and an RBF-kernel SVM¹. Scores are computed using 5 random seeds and 10-fold stratified cross-validation; we report mean accuracy across seeds and folds to remain consistent with the evaluation of downstream GNN baselines. The final SLA and FLA values are defined as the average performance of the LogReg and RBF-SVM probes on the corresponding structural or feature summaries.

Limitations of the current probe models. While this probe suite provides a simple and interpretable baseline, it has several practical limitations. First, the RBF-SVM does not scale favorably with the number of graphs N or with increasing summary density, requiring $\mathcal{O}(N^2)$ memory for

¹After early ablations, we selected these two models to probe both linear separability (LogReg) and simple non-linear separability (RBF-SVM [CV95]) of the summary vectors, while keeping the probes architecture-agnostic and shallow.

the kernel matrix and up to $\mathcal{O}(N^3)$ time in the worst case. As a result, SLA/FLA evaluation can become a computational bottleneck for large graph collections or dense synthetic sweeps. Second, although logistic regression is computationally efficient, it probes only linear decision boundaries and may underestimate alignment when the label signal is weakly non-linear in the summary space. Conversely, the RBF-SVM can overemphasize fine-grained separability that may not translate to realistic graph models, particularly when the summaries are low-dimensional but densely clustered.

Alternative probe suites and alignment planes. Importantly, SLA and FLA are not tied to a specific choice of classifier. The alignment framework can be instantiated with alternative probe suites, thereby inducing different *alignment planes*. For example, replacing the RBF-SVM with cheaper non-linear probes (e.g., random-feature approximations, linear models on polynomial expansions, or small shallow MLPs) yields a computationally lighter alignment evaluation that scales linearly in N . Such variants trade expressivity for efficiency and can be preferable when screening large numbers of datasets or synthetic configurations. Throughout this work, we treat the $(\widehat{\text{SLA}}, \widehat{\text{FLA}})$ plane defined by LogReg+RBF-SVM as a reference instantiation, but emphasize that the framework is modular and admits alternative probe choices depending on computational budget and desired notion of separability.

B.4 Computational Complexity and Classifier Selection

We analyze the computational complexity of the SLA/FLA evaluation pipeline used throughout this work. The procedure consists of two main stages: (i) graph-level summary construction and (ii) repeated classifier evaluation under $s = 5$ random seeds using stratified cross-validation.

Notation. Let N denote the number of graphs in a dataset. For graph i , let $n_i = |V_i|$ and $m_i = |E_i|$ be the number of nodes and edges, respectively. Let F denote the node feature dimension. The number of graph classes is denoted by C . Structural summaries are constructed using $T = 10$ filtration thresholds and $k = 4$ eigenvalues.

Structural Summaries. For each graph, the structural summary includes Betti profiles, spectral features, and standard graph statistics:

- **Betti profiles.** Using a degree-based filtration with $T = 10$ thresholds, each threshold requires scanning all edges and computing connected components, yielding a total cost of

$$\Theta(T(n_i + m_i)).$$

- **Spectral features.** The top $k = 4$ eigenvalues of both the normalized Laplacian and the adjacency matrix are computed using sparse Lanczos methods. Each eigensolve has cost

$$\tilde{\Theta}(k(n_i + m_i)),$$

where $\tilde{\Theta}(\cdot)$ suppresses iteration-dependent constants. Two such eigensolves are performed per graph.

- **Graph statistics.** Degree moments are linear in n_i , while clustering coefficients and triangle counts require

$$\Theta\left(\sum_{v \in V_i} d(v)^2\right),$$

which dominates in graphs with skewed degree distributions.

Combining these terms, the total cost of structural summary construction is

$$\sum_{i=1}^N \left[\Theta(T(n_i + m_i)) + \tilde{\Theta}(k(n_i + m_i)) + \Theta\left(\sum_{v \in V_i} d(v)^2\right) \right].$$

Since T and k are fixed constants, the resulting structural feature dimension is constant with respect to graph size.

Feature summaries. For each graph, the feature summary is computed as the mean of node features, which costs $\Theta(n_i F)$. Summed over the dataset, the total cost is

$$\Theta\left(\sum_{i=1}^N n_i F\right),$$

and the resulting feature matrix has size $N \times F$.

Classifier Evaluation (SLA and FLA). For both SLA and FLA representations, classifiers are evaluated using $s = 5$ random seeds and $K = 10$ stratified folds. Two models are used: logistic regression and an RBF-kernel SVM.

Logistic regression. Training logistic regression using LBFGS requires

$$\Theta(I_{\text{LR}} N d C)$$

per fold, where d is the feature dimension and I_{LR} is the number of optimization iterations. Prediction is linear in Nd .

RBF-kernel SVM. Training an RBF SVM requires between $\tilde{\Theta}(N^2)$ and $\Theta(N^3)$ time per fold, with $\Theta(N^2)$ memory in the worst case due to kernel matrix storage. This step becomes the dominant cost as N grows.

Overall Complexity. Let d_s and d_f denote the dimensions of the structural and feature summaries, respectively. The total time complexity of the SLA/FLA evaluation pipeline is

$$T_{\text{total}} = T_{\text{pre}} + 2sK [\Theta(Nd) + \Theta(I_{\text{LR}} N d C) + T_{\text{SVM}}(N)],$$

where $d \in \{d_s, d_f\}$ and $T_{\text{SVM}}(N) \in [\tilde{\Theta}(N^2), \Theta(N^3)]$. The preprocessing cost T_{pre} is dominated by structural summary construction.

Memory Complexity. Feature storage requires $\Theta(Nd_s + Nd_f)$ memory. During cross-validation, logistic regression requires $\Theta(dC)$ parameters, while the RBF SVM may require up to $\Theta(N^2)$ memory, which is the primary memory bottleneck.

C Probe Robustness and Structural Ablations

This appendix tests how much the alignment plane depends on our design choices: (i) which components of the structural summary vector contribute to SLA; (ii) whether a low SLA or FLA score may be limited by the reference probe suite; (iii) whether the centroid summary biases FLA; and (iv) whether regimes, correlations, and scores hold up when these choices and the random seeds change.

Benchmark sets. The ablations in Sections C.1 and C.3 to C.5 and Table 9 use 19 TU benchmarks: 14 with node attributes and 5 featureless social graphs. FLA and AlignAvg are defined only on the 14. This is not the same set as the 19 FLA-defined benchmarks in Table 1, which also include the OGB molecular and vision datasets but not the social graphs, so correlations reported here are not directly comparable to that table. Re-computed full-descriptor SLA differs from Table 2 by at most 0.35 points.

C.1 Structural Summary Ablations

The structural summary vector $\phi_{\text{struct}}(G) \in \mathbb{R}^{48}$ concatenates three families of permutation-invariant graph descriptors: a degree-filtration Betti profile (PH, 20 dimensions), a compact spectral signature (Spectral, 8 dimensions), and global graph statistics (Stats, 20 dimensions). To assess the contribution of each component, we evaluate SLA using each block in isolation and each pair of blocks, and compare against the full 48-dimensional vector.

Protocol. Each ablation uses the same probe suite and cross-validation protocol as the main experiments (Appendix B.3): 5 seeds, 10-fold stratified cross-validation, and the average of logistic regression and RBF-kernel SVM. The only change across rows is the structural input representation used by the probe.

Table 5: Structural ablation: SLA (%) for each block alone, each pair, and the full 48-dimensional summary vector, on 14 TU benchmarks with node attributes and 5 featureless social benchmarks. **Bold** is the best score per row.

DATASET	PH (20D)	SPECTRAL (8D)	STATS (20D)	PH+SPEC (28D)	PH+STATS (40D)	SPEC+STATS (28D)	FULL (48D)
AIDS	99.39	99.62	99.50	99.59	99.44	99.55	99.53
MUTAG	86.70	87.61	86.50	86.38	86.39	86.13	86.49
DD	73.11	75.37	76.92	75.77	76.55	77.53	76.85
BZR	70.41	77.09	71.32	75.68	75.07	76.86	77.34
PROTEINS_FULL	69.41	73.67	73.32	72.63	72.72	73.83	73.53
PROTEINS	69.41	73.66	73.32	72.57	72.72	73.83	73.48
COX2	46.94	64.78	46.93	59.00	48.70	60.40	61.77
NC11	65.79	64.22	67.17	67.08	67.60	68.01	68.49
FRANKENSTEIN	66.10	62.73	67.26	66.64	67.38	67.99	68.04
SYNTHE	50.18	49.17	48.73	52.15	47.03	50.60	46.95
ENZYMES	28.18	25.25	34.00	31.72	33.95	35.70	35.68
LETTER-HIGH	31.55	36.13	38.40	38.40	37.97	38.76	38.29
COLORS-3	9.83	9.45	9.81	9.37	9.72	9.83	9.60
COIL-DEL	10.02	16.05	16.05	14.35	16.14	18.42	17.76
REDDIT-B	77.02	79.43	85.83	84.64	86.09	88.65	89.00
COLLAB	69.71	66.23	72.41	70.00	73.51	73.07	73.61
IMDB-B	66.23	69.58	71.23	70.19	70.70	71.04	70.55
REDDIT-5k	48.01	48.39	54.41	51.80	54.56	55.21	55.37
IMDB-M	42.70	42.73	46.70	45.71	45.85	46.61	45.83
AVERAGE	56.88	59.01	59.99	60.19	60.11	61.69	61.48

Table 6: Structural ablation summary. Rank correlations are Spearman against the full descriptor, for SLA (19 datasets) and for AlignAvg (the 14 datasets where FLA is defined, with centroid FLA held fixed). Pearson is against GCN accuracy on the 19 datasets. Regime agreement uses per-axis tertiles of the full-descriptor SLA.

BLOCKS	DIM	AVG SLA	SLA ρ	ALIGNAVG ρ	PEARSON (GCN)	REGIME AGR. (%)
PH	20	56.88	0.987	0.982	0.868	73.7
SPECTRAL	8	59.01	0.981	0.996	0.898	78.9
STATS	20	59.99	0.968	0.982	0.840	78.9
PH+SPECTRAL	28	60.19	0.984	0.996	0.887	89.5
PH+STATS	40	60.11	0.994	0.982	0.850	84.2
SPECTRAL+STATS	28	61.69	0.993	0.996	0.882	78.9
FULL	48	61.48	1.000	1.000	0.880	100.0

Findings. Table 5 shows that the best individual component varies by dataset and graph domain. Comparing the three single blocks and the full vector, Spectral is strongest on five datasets, predominantly molecular and biological graphs (AIDS, MUTAG, PROTEINS, PROTEINS-full, COX2). Stats is strongest on four (DD, Letter-High, IMDB-B, IMDB-M). PH leads on Synthie and COLORS-3, though on COLORS-3 all variants sit near chance (≈ 9 – 10% over 11 classes), so structure carries essentially no signal there. The full descriptor is best on the remaining eight. On COX2, Spectral (64.78) outperforms the full vector (61.77) by a notable margin, so combining components can occasionally add noise on small benchmarks. On REDDIT-B, the full vector (89.00) gains the most over any single block (Stats, 85.83).

Table 6 summarizes the effect on rankings. The full vector has a higher average SLA (61.48) than any single block (56.88–59.99), but it is not strictly best: Spectral+Stats, at 28 dimensions and without PH, averages 61.69. Rankings barely depend on the choice. Every configuration keeps SLA rank correlation of at least 0.968 and AlignAvg rank correlation of at least 0.982 with the default. The AlignAvg figures are partly mechanical, since half of AlignAvg (centroid FLA) is held fixed. We keep the full vector as the default because it covers all three families, not because it dominates them.

Minimal descriptor. Table 7 ranks the 48 dimensions by mutual information with the label, averaged across datasets. A 12-dimensional subset with no PH features (all 8 spectral dimensions and 4 statistics) reaches a rank correlation of 0.988 with the full vector. Larger subsets add little to the ranking and do not improve regime agreement. Since selection and evaluation use the same datasets, this is an optimistic estimate and a starting point for a lightweight variant rather than a validated replacement.

Table 7: Minimal structural descriptor: top- k dimensions by mutual information, averaged across datasets. Lap and Adj are the Laplacian and adjacency spectral blocks. Dimensions are selected and evaluated on the same 19 datasets, so these figures are in-sample.

k	BLOCKS SELECTED	AVG SLA	SLA ρ VS. FULL	PEARSON (GCN)	REGIME AGR. (%)
8	3 LAP, 2 ADJ, 3 STATS	59.97	0.972	0.870	68.4
12	4 LAP, 4 ADJ, 4 STATS	60.79	0.988	0.884	78.9
16	4 PH, 4 LAP, 4 ADJ, 4 STATS	60.99	0.988	0.880	78.9
24	8 PH, 4 LAP, 4 ADJ, 8 STATS	61.71	0.989	0.888	78.9

C.2 Probe Sufficiency Self-Check

The reference alignment scores use a deliberately lightweight probe suite: logistic regression and RBF-kernel SVM. This design keeps SLA and FLA cheap, standardized, and independent of graph encoder training. A natural concern is probe underfitting: a low SLA or FLA score may occur because the restricted-view summary contains little accessible marginal label signal, or because the reference probes are not expressive enough to separate the summary vectors. We therefore introduce a simple probe sufficiency self-check.

Alternative probe suite. We evaluate an alternative, more expressive probe suite consisting of five classifiers: logistic regression, linear SVM, RBF-SVM, a shallow MLP with two hidden layers and ReLU activations, and CatBoost. The same 5-seed, 10-fold stratified cross-validation protocol is used. This alternative suite is not intended to replace the reference protocol, but to test whether the dataset placement in the SLA-FLA plane is stable under a richer family of shallow predictors.

Let $\widehat{SLA}_{\text{ref}}$ and $\widehat{SLA}_{\text{alt}}$ denote the reference and alternative-probe SLA scores, respectively. We define the probe-sensitivity gap

$$\delta_{SLA} = \widehat{SLA}_{\text{alt}} - \widehat{SLA}_{\text{ref}},$$

and analogously define $\delta_{FLA} = \widehat{FLA}_{\text{alt}} - \widehat{FLA}_{\text{ref}}$. Positive values indicate that the alternative suite finds more accessible marginal signal than the reference suite, while negative values indicate that the reference suite is at least as strong under this protocol.

Self-check interpretation. We interpret the gap as a probe-sensitivity diagnostic, not as a formal statistical test:

- $|\delta| \leq 5$ points: the reference score is empirically stable under the alternative probe suite, suggesting that probe underfitting is unlikely to be the dominant explanation for the observed alignment score.
- $|\delta| > 5$ points: the reference score should be treated as probe-sensitive. In this case, the alternative-probe score should be reported alongside the reference score, and conclusions about low alignment should be made cautiously.

The 5-point threshold is a practical heuristic based on the observed gap distribution in Table 8. It is intended to flag clear probe-sensitivity cases rather than define a universal cutoff.

Results. Table 8 shows that the reference alignment scores are stable for most datasets. For 15 of the 17 datasets, $|\delta_{SLA}| \leq 5$ points. Gaps in both directions are roughly equally common, confirming that the reference protocol does not systematically underestimate marginal accessibility: negative gaps, where the reference suite matches or exceeds the alternative, are as frequent as positive ones.

The clearest exception is LETTER-HIGH, with $\delta_{SLA} = +11.43$ and $\delta_{FLA} = +21.54$. This dataset is atypical in our suite: it contains small graphs with average size 4.67, 2-dimensional node features, and 15 classes. The large positive gap indicates that the reference probes underestimate marginal accessibility on this dataset, while the alternative suite can exploit additional fine-grained feature geometry. We therefore treat LETTER-HIGH as probe-sensitive rather than genuinely low-alignment under all shallow probes.

Table 9: Individual probes on the same summaries. ρ and regime agreement are measured against LogReg alone, not against the two-probe reference average. SLA uses 19 datasets and FLA uses 14.

PROBE	SLA			FLA		
	ρ	PEARSON (GCN)	REGIME (%)	ρ	PEARSON (GCN)	REGIME (%)
LOGREG	1.000	0.880	100.0	1.000	0.931	100.0
LINEAR SVM	0.989	0.892	84.2	1.000	0.925	100.0
RBF-SVM	0.989	0.877	84.2	0.947	0.930	85.7
MLP	0.935	0.918	73.7	0.925	0.911	78.6
CATBOOST	0.909	0.926	78.9	0.930	0.789	71.4

This interpretation is also consistent with the observed backbone spread on LETTER-HIGH: architectures with more expressive feature and attention mechanisms, such as SAGE and GPS, achieve substantially higher accuracy than GCN. Thus, a large probe gap is itself informative: it flags datasets where marginal signal may exist but requires more expressive access than the reference probe provides.

Practical guidance. The self-check adds one lightweight step to the alignment pipeline. After computing the reference SLA and FLA scores, practitioners can run one or more stronger shallow probes, such as an MLP or CatBoost, on the same restricted-view summaries. If the resulting gap is small, the reference score can be treated as stable. If the gap is large, the dataset should be flagged as probe-sensitive, and both the reference and alternative-probe scores should be reported.

We emphasize that this self-check does not change the interpretation of SLA and FLA. They remain restricted-view diagnostics of marginal label accessibility under specified summary maps and probe families. The self-check simply guards against a known edge case: a low reference score caused by insufficient probe expressivity rather than weak signal in the restricted-view summary.

Table 8: Probe sufficiency self-check: difference between alternative-probe and reference alignment scores. We report $\delta_{\text{SLA}} = \widehat{\text{SLA}}_{\text{alt}} - \widehat{\text{SLA}}_{\text{ref}}$ and analogously for FLA. Most datasets show small, unsystematic differences. LETTER-HIGH is the clearest probe-sensitive case.

DATASET	δ_{SLA}	δ_{FLA}
AIDS	+0.06	-1.24
MOLTOX21	-3.13	-2.92
MUTAG	-1.71	+0.78
DD	-0.13	-3.97
BZR	-7.16	-11.07
PROTEINS_FULL	-2.18	-0.25
MOLHIV	-4.33	-5.53
BACE	-1.57	-4.60
PROTEINS	-2.14	-4.68
COX2	+4.86	-8.26
NCI1	-1.18	-2.32
FRANKENSTEIN	-0.14	-2.59
ENZYMES	-2.29	+5.57
MNIST	-0.10	-0.43
CIFAR-10	+0.35	-1.78
COIL-DEL	-2.10	-4.29
LETTER-HIGH	+11.43	+21.54

Individual probes. Table 9 scores each probe separately on the same summaries. Rank correlation with LogReg stays between 0.909 and 1.000. Pearson correlation with GCN accuracy stays between 0.877 and 0.926 for SLA. For FLA it stays between 0.911 and 0.931, except CatBoost at 0.789, which raises FLA sharply on ENZYMES and COLORS-3 relative to the other probes.

C.3 Feature Summary Robustness

Centroid pooling discards higher-order information about the node-feature distribution, so an under-powered summary could create or enlarge the low-alignment regime. We therefore recompute FLA under four richer permutation-invariant summaries: per-feature mean and standard deviation; mean, standard deviation, and skewness; quantile pooling at 10/50/90%; and the top-4 eigenvalues of the feature covariance. The probe protocol is unchanged.

Findings. Richer summaries do raise FLA: average FLA rises from 54.70 for the centroid to 60.10 for mean+std+skew (Table 10), so the centroid is not the highest-scoring summary. The effect is concentrated on some low-alignment datasets. Under mean+std+skew, COIL-DEL moves from 5.70 to 21.48, Letter-High from 23.27 to 37.36, and ENZYMES from 36.38 to 51.10. Low centroid FLA on these datasets is therefore partly a property of the summary, and we read it as a lower bound on feature accessibility.

Table 10: FLA (%) under five feature summaries on the 14 TU benchmarks with node attributes. **Bold** is the best score per row.

DATASET	CENTROID	MEAN+STD	MEAN+STD+SKEW	QUANTILE (10/50/90)	COV. EIGS (TOP-4)
AIDS	96.86	96.12	97.28	94.74	77.69
MUTAG	69.58	67.58	66.42	67.44	69.69
DD	72.51	77.14	76.58	69.50	65.46
BZR	71.79	75.10	77.00	75.58	64.53
PROTEINS_FULL	74.08	74.38	74.27	74.10	71.46
PROTEINS	64.02	71.81	71.67	70.36	67.88
COX2	67.13	73.75	74.33	71.76	63.02
NCI1	59.81	63.55	64.16	57.80	59.55
FRANKENSTEIN	54.66	55.15	56.66	55.92	55.90
SYNTHIE	45.83	44.10	44.22	41.40	29.38
ENZYMES	36.38	45.12	51.10	43.60	24.68
LETTER-HIGH	23.27	33.50	37.36	34.24	23.46
COLORS-3	24.22	28.65	28.82	22.30	26.50
COIL-DEL	5.70	12.28	21.48	18.01	5.06
AVERAGE	54.70	58.45	60.10	56.91	50.30

Table 11: Feature-summary robustness on 14 datasets. Rank correlations are Spearman against centroid FLA, for FLA and for AlignAvg (with full SLA held fixed). Regime agreement uses per-axis tertiles of centroid FLA.

SUMMARY	DIM	AVG FLA	FLA ρ	ALIGNAVG ρ	PEARSON (GCN)	REGIME AGR. (%)
CENTROID	d	54.70	1.000	1.000	0.934	100.0
MEAN+STD	$2d$	58.45	0.965	1.000	0.920	78.6
MEAN+STD+SKEW	$3d$	60.10	0.943	0.978	0.925	71.4
QUANTILE (10/50/90)	$3d$	56.91	0.934	0.982	0.936	78.6
COV. EIGS (TOP-4)	4	50.30	0.947	0.978	0.842	85.7

The ordering and the relation to accuracy are nevertheless stable (Table 11). Rank correlation with centroid FLA is at least 0.934, and AlignAvg rank correlation is at least 0.978. Pearson correlation with GCN accuracy is 0.920–0.936 for mean+std, mean+std+skew, and quantile pooling, against 0.934 for the centroid; it falls to 0.842 for covariance eigenvalues, which discard the mean entirely. We keep the centroid as the default for its cost and simplicity. We do not evaluate trained summaries such as Deep Sets, because a summary fitted to labels would reintroduce the encoder training the diagnostic is designed to avoid.

C.4 Regime and Correlation Stability

We now test whether the central conclusions, the low-alignment regime and the backbone-agnostic trend, survive these changes jointly. Table 12 assigns each dataset a joint regime using fixed cutoffs and reports how many datasets change regime under each construction, together with the AlignAvg–accuracy correlation for all four backbones. The cutoffs are a reporting convention, since regime labels are relative to the chart that defines them.

Findings. No dataset jumps directly between high/high and low/low under any construction. Every change is a single step into or out of the mixed regime, and at most 5 of 17 datasets change under any one construction. Labels near the cutoffs do move: ENZYMES and COX2 change most often. The per-axis tertile regimes in Tables 6, 9 and 11 show the same behaviour with no direct High–Low jumps, and agreement of 71–90% for non-reference constructions reflects moves to and from the middle tier.

The backbone trend also holds. Across the 14-dataset rows, correlations stay within 0.898–0.989 for GCN and GAT and within 0.611–0.817 for SAGE and GPS, so the weaker relationship for SAGE and GPS is itself stable. In the five-probe plane on 17 datasets, correlations move from 0.955/0.764/0.957/0.789 to 0.925/0.812/0.916/0.849 (GCN/SAGE/GAT/GPS). The shifts run in both directions and neither plane is uniformly better. Figure 4 shows that the layout of the plane is visually preserved under alternative constructions.

Table 12: Stability of joint regimes and of the backbone trend. A dataset is high/high if SLA and FLA are both ≥ 60 , low/low if both are < 40 , and mixed otherwise. **Changed** counts datasets whose joint regime differs from the reference row of the same block. **HH $\leftrightarrow$ LL** counts direct jumps between the two extremes. The last four columns are Pearson correlations between AlignAvg and accuracy. All rows above the last block use the 14 TU benchmarks with node attributes.

CONSTRUCTION	CHANGED	HH $\leftrightarrow$ LL	GCN	SAGE	GAT	GPS
REFERENCE (FULL SLA, CENTROID FLA)	0/14	0	0.939	0.709	0.970	0.775
FLA: MEAN+STD	2/14	0	0.935	0.715	0.973	0.799
FLA: MEAN+STD+SKEW	2/14	0	0.937	0.728	0.973	0.814
FLA: QUANTILE (10/50/90)	1/14	0	0.938	0.735	0.972	0.817
FLA: COV. EIGS (TOP-4)	0/14	0	0.898	0.663	0.955	0.767
SLA: PH	1/14	0	0.928	0.692	0.952	0.740
SLA: SPECTRAL	0/14	0	0.944	0.719	0.968	0.772
SLA: STATS	1/14	0	0.924	0.699	0.950	0.758
SLA: PH+SPECTRAL	1/14	0	0.942	0.717	0.964	0.767
SLA: PH+STATS	1/14	0	0.926	0.700	0.956	0.763
SLA: SPECTRAL+STATS	0/14	0	0.941	0.714	0.967	0.770
PROBE: LOGREG	2/14	0	0.941	0.711	0.969	0.773
PROBE: LINEAR SVM	2/14	0	0.945	0.706	0.973	0.769
PROBE: RBF-SVM	2/14	0	0.935	0.707	0.969	0.775
PROBE: MLP	2/14	0	0.949	0.693	0.989	0.768
PROBE: CATBOOST	3/14	0	0.919	0.611	0.978	0.713
FIVE-PROBE SUITE (TABLE 8), 17 DATASETS, REFERENCE	0/17	0	0.955	0.764	0.957	0.789
FIVE-PROBE SUITE (TABLE 8), 17 DATASETS, ALTERNATIVE	5/17	0	0.925	0.812	0.916	0.849

Table 13: Stability across 5 seeds, each drawing new 10-fold splits and refitting the probes. GCN accuracies are held fixed at their Table 2 values. The last two columns are the mean and largest per-dataset standard deviation of the score, in accuracy points.

VIEW	DATASETS	PEARSON (GCN)	SPEARMAN (GCN)	MEAN STD	MAX STD
SLA	19	0.880 $\pm$ 0.002	0.842 $\pm$ 0.012	0.33	1.07
FLA	14	0.934 $\pm$ 0.003	0.889 $\pm$ 0.008	0.45	1.15

C.5 Seed and Split Stability

We recompute SLA and FLA under 5 seeds, each drawing new 10-fold stratified splits and refitting the probes (Table 13). Correlations with GCN accuracy barely move, and the average per-dataset standard deviation across both views is 0.38 accuracy points, with a maximum of 1.15 (FLA on Synthie). This test varies the alignment scores only, since GCN accuracies are held fixed. It also covers random in-distribution splits only; Section H.4 examines scaffold splits, where the picture differs.

D Heterophilic Ego-Networks

Heterophilic benchmarks test a different failure mode than the synthetic Twins: rather than suppressing marginal shortcuts by construction, they suppress them through label-structure anticorrelation. We convert four standard heterophilic node-classification graphs—Texas, Cornell, Wisconsin (WebKB), and Actor—into graph classification tasks by extracting k -hop ego-subgraphs and using the center-node label as the graph label, with a center-node indicator feature included in all experiments. We vary $k \in \{1, 2, 3\}$ to probe how much higher-order structure is needed before alignment and accuracy become recoverable. Dataset statistics are in Table 14.

Why heterophilic datasets require separate treatment. In heterophilic regimes, immediate neighborhoods are label-adversarial by construction: 1-hop aggregation actively mixes signal from nodes of different classes. Increasing k exposes longer-range structure that may restore label signal, but at the cost of incorporating more label-misaligned context. This bias–variance tradeoff is fundamentally different from the homophilic case and makes heterophilic ego-graphs a stress test for whether alignment scores track accuracy when the source of label signal is relational rather than marginal.

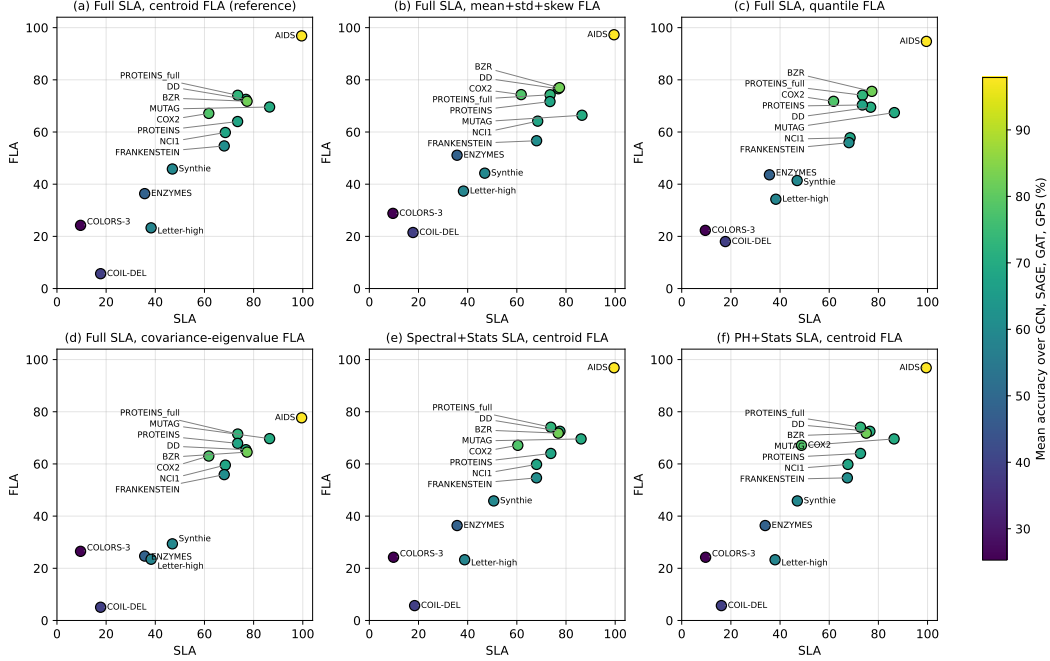


Figure 4: **The SLA–FLA plane under alternative constructions.** The 14 TU benchmarks with node attributes, colored by mean accuracy over the four backbones. Panel (a) is the reference construction; panels (b)–(d) change the feature summary and panels (e)–(f) change the structural blocks.

Table 14: Dataset statistics for Node-to-Graph heterophilic ego-graph benchmarks at varying k -hop sizes.

DATASET	k	#GRAPHS	AV.NODES	AV.EDGES	# FEAT.	#CLASSES
TEXAS	1	183	4.05	9.83	1704	5
TEXAS	2	183	66.68	215.73	1704	5
TEXAS	3	183	127.83	450.69	1704	5
CORNELL	1	183	4.03	8.74	1704	5
CORNELL	2	183	55.86	146.09	1704	5
CORNELL	3	183	116.04	360.03	1704	5
WISCONSIN	1	251	4.59	11.82	1704	5
WISCONSIN	2	251	69.76	245.19	1704	5
WISCONSIN	3	251	158.31	637.61	1704	5
ACTOR	1	7600	7.88	22.23	933	5
ACTOR	2	7600	121.82	374.78	933	5
ACTOR	3	7600	227.979	379.08	933	5

D.1 Results

Results are summarized in Table 15. The headline finding is that all pooled message-passing baselines achieve low accuracy across all datasets and hop sizes (typically 22–48%, only modestly above 20% chance for five classes), consistent with the low-alignment characterization. However, two patterns within this overall failure are worth isolating.

SLA rises with k but accuracy does not follow. Across all three WebKB datasets, structural alignment increases monotonically with hop radius—Texas rises from $SLA = 33.78$ at $k = 1$ to 62.67 at $k = 3$, with Cornell and Wisconsin showing smaller but consistent gains. FLA remains stable throughout, confirming that the additional signal is structural rather than feature-driven. Yet accuracy peaks at $k = 1$ or $k = 2$ for most backbones and degrades at $k = 3$. This dissociation

Table 15: Pooled-model accuracy (%) tracked against the alignment plane for heterophilic ego-graph datasets at varying hop sizes k . AlignAvg summarizes alignment as $(\text{SLA} + \text{FLA})/2$.

DATASET	k	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
TEXAS	1	31.05±10.84	42.11±6.86	35.79±15.49	42.11±6.86	33.78	43.16	38.47
TEXAS	2	35.26±4.88	36.32±6.32	25.79±13.87	36.32±6.09	51.81	61.96	56.89
TEXAS	3	30.00±4.59	35.79±7.37	22.63±12.74	37.89±4.88	62.67	61.84	62.26
CORNELL	1	36.32±3.87	35.26±12.85	35.26±6.36	60.00±7.88	17.76	31.34	24.55
CORNELL	2	27.89±4.88	25.79±1.97	22.11±3.57	58.95±2.68	29.13	32.29	30.71
CORNELL	3	27.37±6.78	33.68±9.47	20.53±3.07	58.95±1.29	28.89	32.18	30.54
WISCONSIN	1	39.22±3.28	41.57±4.54	32.16±4.90	52.16±6.97	27.86	44.72	36.29
WISCONSIN	2	48.24±10.49	37.65±8.36	33.33±6.44	53.73±2.00	37.99	45.45	41.72
WISCONSIN	3	39.61±5.02	41.96±10.27	34.90±8.72	50.98±6.20	34.93	46.08	40.51
ACTOR	1	23.37±0.45	27.30±0.69	21.89±0.49	27.20±1.14	18.20	24.12	21.16
ACTOR	2	22.76±0.73	25.08±0.18	19.99±0.74	25.75±0.36	19.81	20.05	19.93
ACTOR	3	20.80±1.03	23.95±1.23	18.71±1.57	26.22±1.29	9.99	19.87	14.93

between alignment and accuracy reveals a limitation of SLA as a predictor in heterophilic settings: our structural probe is permutation-invariant and captures global topology, but the discriminative signal at larger k is increasingly entangled with label-misaligned context that pooled readouts cannot filter. Higher alignment does not help when the model cannot exploit it.

GPS is anomalously strong at small k . On Cornell at $k = 1$, GPS achieves 60.0% accuracy despite very low alignment ($\text{SLA} = 17.76$, $\text{FLA} = 31.34$)—a 25-point gap over GCN and a clear outlier relative to the alignment-prediction story. This is not noise: the same pattern holds on Texas and Wisconsin at $k = 1$. The likely explanation is that GPS’s global attention and positional encodings allow it to exploit center-node identity and sparse structural cues that local MPNN aggregation cannot access when neighborhoods are adversarial. This constitutes a genuine scope boundary for the GLA predictor: in heterophilic settings at small k , architecture choice matters beyond what alignment scores capture, and GPS is the appropriate default.

Actor: alignment destruction under ego expansion. Actor is the cleanest case in this section. SLA and FLA are uniformly low across all hop sizes (AlignAvg drops from 21.16 at $k = 1$ to 14.93 at $k = 3$), and accuracy decreases monotonically with k for every backbone except GPS. Ego-graph expansion does not recover signal here—it destroys it, by diluting discriminative node features with structurally irrelevant context. Actor is a feature-dominated heterophilic regime where the center node’s features carry the label signal and any aggregation beyond $k = 1$ makes the task strictly harder.

Summary. Heterophilic ego-graphs occupy a regime where the standard alignment–accuracy relationship breaks down in two distinct ways: SLA can rise while accuracy falls (WebKB, large k), and GPS can outperform its alignment score by a large margin (WebKB, small k). Both patterns point to the same underlying issue—pooled readouts cannot selectively aggregate label-aligned structure when label-misaligned context dominates the ego-graph. The practical implication is concrete: in heterophilic node-to-graph transfer, alignment scores remain useful for ruling out hopeless configurations (Actor at $k = 3$) but are insufficient as predictors of achievable accuracy; backbone choice and ego-graph radius must be tuned jointly.

E Synthetic Sweep Experiments

This appendix provides full implementation details for the synthetic alignment experiments described in Section 4. Its purpose is to ensure reproducibility, clarify how alignment scores behave under controlled interventions, and formally justify the diagnostic role of the Twins construction.

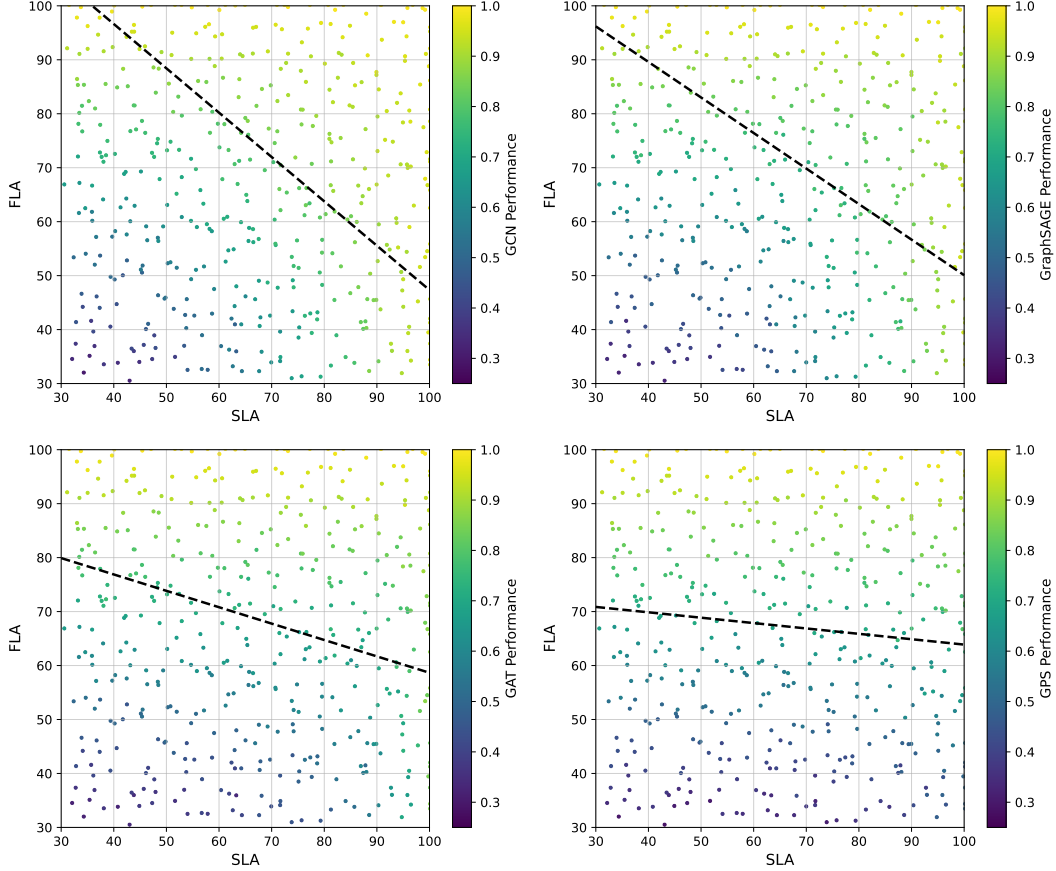


Figure 5: **GNN performance across synthetic sweep with boundaries.** For a full synthetic sweep across the SLA-FLA plane colored by GNN accuracy. The dashed line marks a decision boundary for “good” performance set at 80% accuracy. We include results for all four backbones, beginning in the top left: GCN, GraphSAGE, GAT, and GPS.

E.1 Synthetic Dataset Generation Details

All synthetic datasets are generated with fixed global statistics unless explicitly varied by sweep parameters. Each dataset consists of $N = 1000$ graphs per sweep point. Graph sizes are sampled independently as $n \sim \mathcal{N}(120, 20^2)$ and clipped to $[60, 200]$. Node feature dimension is fixed to $d = 16$.

For alignment plane sweeps, the structural separation parameter $s \in [0, 1]$ and feature separation parameter $f \in [0, 1]$ are discretized on a uniform 15×15 grid. Each grid point corresponds to an independently generated dataset, with identical random seeds reused across all GNN backbones.

Unless otherwise stated, Gaussian noise variance is fixed to $\sigma^2 = 0.1$. Class mean vectors $\{\mu_y\}$ are sampled once per dataset family, normalized to unit norm, and scaled to a fixed signal magnitude.

E.2 Structural Signal Control and SLA Behavior

Structural label alignment (SLA) is controlled exclusively by the structural separation parameter s . When $s = 0$, all graphs are sampled from the shared generator $\mathcal{G}_0$, yielding identical structural distributions across classes and SLA near chance. As $s \rightarrow 1$, graphs are increasingly sampled from class-specific generators $\mathcal{G}_y$, which differ in global structural statistics such as degree distributions, clustering coefficients, and mesoscopic community structure.

Algorithm 1 Twin synthetic graph generation (one graph)

- 1: Sample SBM graph $G = (V, E)$ with fixed (K, P) and $|V| = n$.
 - 2: Compute node scores $s(v)$ using a chosen filtration (Linear, Radial, PageRank, Eigen).
 - 3: Threshold $s(v)$ at empirical quantiles to form three groups V_1, V_2, V_3 (roughly balanced).
 - 4: Assign features: for $v \in V_j$, set $x_v = \mu_j + \epsilon_v$, $\epsilon_v \sim \mathcal{N}(0, \sigma^2 I)$.
 - 5: Output (G, X) with class label determined by the assignment rule(s) (random vs. structured).
-

Graph sizes and expected degrees are matched across classes, ensuring that increases in SLA are not driven by trivial size or density cues. Empirically, measured $\widehat{\text{SLA}}$ increases monotonically with s across all sweep families, validating s as a calibrated control knob for structural label signal.

E.3 Feature Signal Control and FLA Behavior

Feature label alignment (FLA) is controlled by the feature separation parameter f . When $f = 0$, all node features are i.i.d. Gaussian noise, matched across labels, and pooled feature summaries are non-informative. As f increases, a class-dependent mean shift μ_y is added to all nodes in graphs of label y , causing graph-level feature aggregates to become increasingly predictive.

Noise variance and feature dimensionality are fixed across all sweep points, so that $\widehat{\text{FLA}}$ varies smoothly and monotonically with f . This construction isolates marginal feature signal: no structural information is required to achieve high FLA.

E.4 Alignment Plane Coverage

Jointly sweeping (s, f) induces datasets spanning the full alignment plane, including high-SLA/low-FLA, low-SLA/high-FLA, and jointly aligned regimes. Measured alignment scores $(\widehat{\text{SLA}}, \widehat{\text{FLA}})$ closely track the intended control parameters (s, f) , with small deviations attributable to finite-sample effects and injected noise.

We verify that sweep families occupy well-separated regions of the alignment plane and that no unintended correlations between SLA and FLA arise outside the jointly aligned regime. This confirms that the plane sweep provides a controlled testbed for analyzing how model performance organizes with respect to alignment.

All synthetic sweep output can be viewed in Table 18 and Figure 5.

E.5 Twins Construction and Shortcut Elimination

The Twins construction is designed to eliminate both structure-only and feature-only shortcuts. All classes share: (i) identical SBM parameters and identical edge sets per base graph, and (ii) identical multisets of node features.

As a result, classifiers operating solely on graph structure or on pooled feature statistics achieve near-chance performance. Labels differ only in the coupling between node roles and node feature types, yielding a strictly interaction-only regime.

E.6 Role Scores and Balanced Assignments

For each base graph, nodes are assigned a scalar score $s(v)$ using one of four filtrations: **Linear**, **Radial** [Sei83], **PageRank** [PBMW99], or **Eigen**. The scores are then partitioned into three equal-sized groups according to the first and second tertiles of the score distribution. Feature prototypes (μ_1, μ_2, μ_3) are assigned deterministically to these groups, with additive Gaussian noise. This guarantees that all classes have identical prototype counts, ensuring that marginal feature distributions are exactly matched.

The algorithm for each of the scalar scoring functions is described below.

Linear. Nodes are scored by shortest-path distance from a peripheral root. A degree-1 root is chosen when available, else the next smallest-degree node is used. This permutation yields distance-based stratification, with distinct feature prototypes occupying successive shortest-path shells.

Radial. Nodes are scored by core number, defined as the largest k such that the node belongs to a k -core, a maximal subgraph in which every node is of degree k or greater. This induces core-based feature stratification, with the different feature prototypes occupying successive shells radiating outward from the densest regions of the graph.

PageRank. Nodes are scored by PageRank, which computes the stationary probability of a random walker occupying a given node under a teleportation model. This produces global-importance stratification, with higher-scoring nodes tending to lie at well-connected positions that link multiple regions of the graph.

Eigen. Nodes are scored by eigenvector centrality, which is calculated for each connected component and linearly normalized to $[0, 1]$, with singletons assigned a zero score. This induces spectral-importance stratification in which high-score nodes connect to other high-score nodes, forming a per-component hub-periphery structure. This permutation scheme is similar to the radial distribution when dense cores are well connected, but eigenvector centrality captures a more global influence, whereas k -core only measures local subgraph density.

E.7 Interpretation of Alignment Scores in Twins

Although Twins datasets exhibit low measured SLA and FLA, this does not contradict lower-bound intuitions derived from marginal alignment. Both alignment scores are intentionally suppressed by construction; however, labels remain predictable through higher-order interactions between structure and features.

Alignment scores measure marginal separability under pooled summaries. Twins instead isolate a regime in which prediction depends on preserving global correspondence between node roles and feature types. This highlights a fundamental limitation of marginal alignment metrics and motivates evaluation beyond SLA and FLA alone.

E.8 Binary and Multiclass Twins Settings

We evaluate both binary Twins (random vs. one structured assignment) and multiclass Twins (random vs. two structured rules, e.g., Linear vs. Radial). In multiclass settings, chance accuracy decreases accordingly, and successful prediction requires distinguishing multiple interaction patterns despite identical marginal distributions.

F Background on Topological and Spectral Signatures

F.1 Persistent Homology

Persistent Homology (PH) is a central tool in Topological Data Analysis (TDA) that captures multiscale topological features such as connected components, loops, and voids that persist across different scales [DW22]. Originally developed for point cloud data, PH has since been successfully extended to graphs, images, and other data modalities [CA24]. In graph learning, PH provides latent structural insights that complement conventional methods and GNNs, enabling a richer understanding of both local and global graph structure.

PH typically consists of three main steps: constructing a filtration, obtaining persistence diagrams, and performing vectorization. The process begins with *filtration*, where a nested sequence of simplicial complexes is built to track the evolution of topological features. Given an unweighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a filtration function $f : \mathcal{V} \rightarrow \mathbb{R}$ with thresholds $\mathcal{I} = \{\alpha_i\}$, we construct subgraphs $\mathcal{G}^i = (\mathcal{V}_i, \mathcal{E}_i)$, where $\mathcal{V}_i = \{v \in \mathcal{V} \mid f(v) \leq \alpha_i\}$ (See Figure 6 for toy example). The filtration function can be based on graph structural properties (e.g., degree, betweenness centrality) or derived from domain-specific information (e.g., atomic number in molecular graphs). For weighted graphs, edge weights can also serve as a natural filtration [CA24]. The primary goal is to establish a hierarchy among nodes or edges that yields a meaningful sequence of subgraphs aligned with the downstream task. In temporal graphs, a natural choice would be timestamps on edges as a valuable filtration function.

Once the subgraphs are defined, each $\mathcal{G}^i$ is extended to its clique complex $\widehat{\mathcal{G}}^i$, producing a sequence $\widehat{\mathcal{G}}^1 \subseteq \widehat{\mathcal{G}}^2 \subseteq \dots \subseteq \widehat{\mathcal{G}}^N$. As the filtration evolves, topological features such as connected components (0-holes), loops (1-holes), and cavities (2-holes) emerge and disappear. Each k -dimensional feature σ is associated with a birth index b_σ and death index d_σ ($1 \leq b_\sigma < d_\sigma \leq N$), recorded as a point (b_σ, d_σ) in the k -th persistence diagram $\text{PD}_k(\mathcal{G}) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\widehat{\mathcal{G}}^i)\}$. Since persistence diagrams are sets of points in $\mathbb{R}^2$, a *vectorization* step is necessary for compatibility with machine learning models. Techniques such as persistence images, landscapes, silhouettes, and Betti curves [A+23] are commonly used to transform diagrams into fixed-size feature representations.

While PH effectively captures complex structural patterns, its integration into temporal graph learning remains challenging. Most prior methods are limited to static graphs and fail to model evolving topological features. Additionally, combining PH summaries with temporal dynamics and node attributes is nontrivial. To address these issues, we propose a hybrid framework that extracts topological descriptors from temporal subgraphs, aligning them with sliding windows to track structural changes over time. This enables scalable, interpretable temporal graph classification without relying on rigid snapshots or heavy discretization.

F.2 Spectral Signatures for Graphs

Spectral graph theory provides permutation-invariant summaries of graph structure through eigenvalues of matrices such as the adjacency matrix A and the normalized Laplacian $L = I - D^{-1/2}AD^{-1/2}$ [Chu97, Spi11]. Since eigenvalues are invariant under node relabeling, they offer compact, canonical descriptors that reflect global connectivity patterns, expansion, and community structure.

In this work, we use a *truncated spectrum* rather than a full density-of-states histogram. Concretely, our spectral vector concatenates (i) the four smallest nonzero eigenvalues of L and (ii) the four largest eigenvalues of A (see Appendix Section B.1). The low end of the Laplacian spectrum captures coarse connectivity and bottlenecks: the multiplicity of the 0 eigenvalue equals the number of connected components, and the first nonzero eigenvalue (algebraic connectivity) quantifies how well-knit the graph is, with larger values indicating fewer global bottlenecks [Chu97]. The top adjacency eigenvalues capture dominant interaction scales and degree heterogeneity; for many graph families they correlate with the presence of hubs and overall edge concentration, providing a complementary view to Laplacian features [Chu97].

Using only a small number of eigenvalues yields a fixed-budget, low-dimensional summary that is easy to compute and compare across graphs of varying size, while still encoding informative global structure. This design matches our goal in constructing lightweight structural probes: the truncated spectral signature adds global structural sensitivity beyond local statistics without introducing high-dimensional histograms or model-specific embeddings.

G Per-Backbone Graph Label Alignment

SLA and FLA scores are backbone-independent, using our synthetic sweep across the alignment atlas we can then target backbone-specific recipes to predict GNN performance by fitting a linear predictor

$$\text{GLA} = \alpha \cdot \text{SLA} + (1 - \alpha) \cdot \text{FLA}$$

The optimal convex combinations across all backbones are

$$\begin{aligned} \text{GLA}_{\text{GCN}} &= 0.45\text{SLA} + 0.55\text{FLA}, & \text{GLA}_{\text{SAGE}} &= 0.40\text{SLA} + 0.60\text{FLA}, \\ \text{GLA}_{\text{GAT}} &= 0.23\text{SLA} + 0.77\text{FLA}, & \text{GLA}_{\text{GPS}} &= 0.09\text{SLA} + 0.91\text{FLA}. \end{aligned}$$

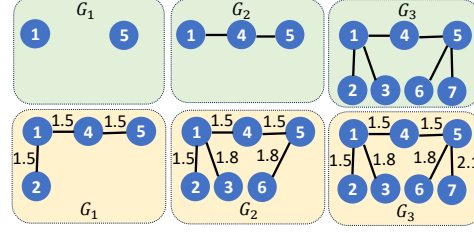


Figure 6: Graph Filtration: For $\mathcal{G} = \mathcal{G}_3$ in both examples, the top figure illustrates a *superlevel filtration* using the *node degree function* with thresholds $3 > 2 > 1$, where nodes of degree 3 are activated first, followed by those of lower degrees. Similarly, the bottom figure illustrates a *sublevel filtration* based on *edge weights* with thresholds $1.5 < 1.8 < 2.1$.

Table 16: Absolute deviation between predicted GLA scores and observed pooled accuracy. For each backbone, $\Delta_{\text{BB}} = |\widehat{\text{GLA}}_{\text{BB}} - \text{Acc}_{\text{BB}}|$. Accuracies use reported means. Featureless datasets (FLA undefined) are omitted.

DATASET	$\widehat{\text{GLA}}_{\text{GCN}}$	GCN	Δ_{GCN}	$\widehat{\text{GLA}}_{\text{SAGE}}$	SAGE	Δ_{SAGE}	$\widehat{\text{GLA}}_{\text{GAT}}$	GAT	Δ_{GAT}	$\widehat{\text{GLA}}_{\text{GPS}}$	GPS	Δ_{GPS}
AIDS	98.05	98.40	0.35	97.91	98.30	0.39	97.46	95.95	1.51	97.09	98.85	1.76
MOLTOX21	86.75	73.45	13.30	86.97	73.63	13.34	87.71	71.95	15.76	88.32	73.70	14.62
MUTAG	77.32	71.28	6.04	76.47	69.74	6.73	73.57	67.18	6.39	71.18	70.77	0.41
DD	74.41	70.30	4.11	74.19	70.55	3.64	73.45	69.03	4.42	72.84	71.90	0.94
BZR	74.17	81.48	7.31	73.89	80.00	6.11	72.93	83.46	10.53	72.14	83.21	11.07
PROTEINS_FULL	73.80	71.07	2.73	73.82	71.70	2.12	73.89	73.84	0.05	73.95	74.11	0.16
MOLHIV	69.61	71.41	1.80	68.97	74.47	5.50	66.79	73.39	6.60	64.98	71.26	6.28
BACE	68.60	72.76	4.16	68.35	77.17	8.82	67.49	60.53	6.96	66.78	69.74	2.96
PROTEINS	68.28	67.59	0.69	67.80	68.12	0.32	66.17	67.05	0.88	64.85	68.93	4.08
COX2	64.61	78.09	13.48	64.86	78.30	13.44	65.73	76.81	11.08	66.45	77.87	11.42
NCI1	63.66	68.30	4.64	63.23	69.39	6.16	61.78	64.53	2.75	60.53	69.39	8.86
FRANKENSTEIN	60.66	59.12	1.54	59.99	62.07	2.08	57.73	61.57	3.84	55.86	60.53	4.67
SYNTHIE	46.57	69.00	22.43	46.51	77.00	30.49	46.28	45.50	0.78	46.10	45.00	1.10
ENZYMES	36.09	47.50	11.41	36.14	50.33	14.19	36.31	43.00	6.69	36.44	50.67	14.23
LETTER-HIGH	29.96	41.38	11.42	29.21	79.29	50.08	26.65	30.80	4.15	24.36	81.51	57.15
MNIST	24.77	37.85	13.08	24.67	48.80	24.13	24.34	32.01	7.67	24.07	47.33	23.26
CIFAR-10	21.49	27.77	6.28	21.77	39.85	18.08	22.71	31.70	8.99	23.48	41.54	18.06
COLORS-3	17.65	24.92	7.27	18.38	26.08	7.70	20.84	25.08	4.24	22.87	25.03	2.16
COIL-DEL	11.10	33.97	22.87	10.50	55.59	45.09	8.44	18.85	10.41	6.75	46.08	39.33

G.1 Evaluation On Real Datasets

Predicted performance, along with computed absolute differences, can be viewed in Table 16.

Across **molecular benchmarks** [MKB⁺20], deviations are consistently small, particularly for GAT and GPS, indicating that classification accuracy in this regime is well explained by alignment alone. These datasets are characterized by moderate graph sizes, low class counts (predominantly binary), and semantically meaningful node features, placing them squarely within the regime where SLA and FLA act as sufficient summary statistics of task difficulty. In this setting, GLA functions as a faithful, backbone-aware predictor of performance.

In contrast, **vision-style graph datasets** (e.g., MNIST, CIFAR-10, Letter-High, COIL-DEL) exhibit large and systematic deviations, with GLA substantially under-predicting observed accuracy, especially for feature-centric backbones. These datasets combine weak or artificially induced graph topology with low-dimensional but highly informative node features and higher class counts. Consequently, performance is dominated by raw feature geometry rather than graph-level alignment, violating the assumptions underlying GLA and revealing a clear failure mode of alignment-based predictors outside alignment-limited regimes.

The **synthetic benchmarks** provide controlled validation of backbone-specific inductive biases. In these settings, attention-based and global models (GAT and GPS) closely track GLA predictions, while classical message-passing architectures (GCN and GraphSAGE) exhibit larger deviations. This behavior aligns with the designed disentanglement of structural and feature cues in the synthetic data and confirms that the learned GLA weights meaningfully capture differences in backbone reliance on feature versus structural alignment.

Finally, for **social graph benchmarks**, node features are absent and FLA is undefined, rendering GLA inapplicable by construction. Performance in this regime is governed by higher-order structural statistics, graph size effects, and dataset-specific biases that are not captured by SLA alone. This delineates a clear boundary of applicability for alignment-based predictors and motivates the need for richer structural descriptors in featureless settings.

Overall, this analysis establishes GLA as a *domain-conditional predictor* of GNN performance: effective in alignment-limited regimes with informative node features, but insufficient in domains where task signal arises from non-alignment cues or higher-order structure.

Table 17: Practical backbone choice as a function of SLA/FLA regime.

SLA	FLA	Recommended backbone	Rationale
High	High	GCN/SAGE (cheapest)	Alignment is easy; capacity adds little, prefer efficiency.
High	Low/Med	GCN (or SAGE)	Structure-dominant; simpler MPNNs suffice and generalize well.
Med	Med	SAGE/GAT	Balanced regime; moderate capacity often best.
Low/Med	High	GPS (or GAT)	Feature-dominant; attention/global aggregation helps.
Low	Low	GPS/GAT (expect low)	Neither cue aligns; need capacity but accuracy remains limited.

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-001	33.83 $\pm$ 3.21	34.67 $\pm$ 0.29	32.17 $\pm$ 4.07	34.67 $\pm$ 3.06	32.17	34.68	33.42
SYNTH-002	30.33 $\pm$ 3.21	32.83 $\pm$ 3.75	32.17 $\pm$ 5.39	29.50 $\pm$ 2.29	34.37	32.17	33.27
SYNTH-003	42.00 $\pm$ 2.29	39.00 $\pm$ 3.28	41.67 $\pm$ 2.75	37.00 $\pm$ 3.04	36.18	39.75	37.97
SYNTH-004	37.33 $\pm$ 4.48	35.00 $\pm$ 2.00	36.67 $\pm$ 2.52	38.83 $\pm$ 4.54	32.83	37.50	35.17
SYNTH-005	45.00 $\pm$ 0.87	45.33 $\pm$ 3.33	43.17 $\pm$ 2.57	43.00 $\pm$ 6.14	32.90	41.48	37.19
SYNTH-006	42.00 $\pm$ 5.89	42.50 $\pm$ 3.00	41.00 $\pm$ 4.27	43.83 $\pm$ 1.89	34.17	44.30	39.23
SYNTH-007	47.17 $\pm$ 4.65	44.83 $\pm$ 3.75	45.83 $\pm$ 3.06	43.83 $\pm$ 4.01	33.95	46.72	40.33
SYNTH-008	45.33 $\pm$ 1.61	44.67 $\pm$ 1.61	45.50 $\pm$ 1.00	43.50 $\pm$ 4.58	36.80	46.22	41.51
SYNTH-009	50.83 $\pm$ 2.75	52.17 $\pm$ 2.36	50.17 $\pm$ 2.02	53.00 $\pm$ 1.73	32.45	53.43	42.94
SYNTH-010	52.83 $\pm$ 2.52	54.33 $\pm$ 0.29	52.50 $\pm$ 2.18	57.50 $\pm$ 2.18	34.73	54.10	44.42
SYNTH-011	57.50 $\pm$ 4.50	57.17 $\pm$ 2.75	60.00 $\pm$ 4.36	58.00 $\pm$ 3.00	35.98	60.18	48.08
SYNTH-012	53.17 $\pm$ 4.54	54.00 $\pm$ 3.77	55.33 $\pm$ 3.33	57.33 $\pm$ 1.61	39.77	57.25	48.51
SYNTH-013	60.17 $\pm$ 3.79	60.33 $\pm$ 3.75	59.00 $\pm$ 1.73	62.00 $\pm$ 3.77	33.45	61.62	47.53
SYNTH-014	65.50 $\pm$ 4.33	65.83 $\pm$ 1.61	63.33 $\pm$ 2.52	64.50 $\pm$ 4.77	34.65	62.50	48.58
SYNTH-015	64.50 $\pm$ 4.82	65.50 $\pm$ 4.09	63.67 $\pm$ 4.16	64.50 $\pm$ 3.12	33.55	67.17	50.36
SYNTH-016	66.17 $\pm$ 3.06	66.17 $\pm$ 2.02	66.00 $\pm$ 1.80	68.00 $\pm$ 2.18	30.68	66.88	48.78
SYNTH-017	73.33 $\pm$ 2.93	74.17 $\pm$ 1.04	72.67 $\pm$ 1.61	72.17 $\pm$ 3.25	37.43	72.78	55.11
SYNTH-018	69.33 $\pm$ 3.79	72.17 $\pm$ 3.88	69.50 $\pm$ 2.65	70.50 $\pm$ 4.44	33.27	72.03	52.65
SYNTH-019	77.17 $\pm$ 2.25	78.00 $\pm$ 2.65	77.33 $\pm$ 0.76	77.33 $\pm$ 1.53	34.08	76.67	55.38
SYNTH-020	77.33 $\pm$ 3.40	79.33 $\pm$ 4.48	79.50 $\pm$ 6.24	78.17 $\pm$ 5.53	33.35	78.07	55.71
SYNTH-021	79.83 $\pm$ 0.58	80.83 $\pm$ 1.26	79.50 $\pm$ 1.80	80.67 $\pm$ 0.76	33.52	80.08	56.80
SYNTH-022	79.17 $\pm$ 3.21	78.83 $\pm$ 2.84	79.83 $\pm$ 2.57	79.67 $\pm$ 2.75	34.03	81.40	57.72
SYNTH-023	88.67 $\pm$ 2.02	89.83 $\pm$ 2.08	89.17 $\pm$ 2.75	89.83 $\pm$ 0.29	33.10	86.33	59.72
SYNTH-024	84.83 $\pm$ 2.02	85.50 $\pm$ 1.00	85.00 $\pm$ 1.80	84.83 $\pm$ 1.61	33.25	85.23	59.24
SYNTH-025	91.83 $\pm$ 1.04	92.33 $\pm$ 1.15	92.67 $\pm$ 1.53	93.00 $\pm$ 1.73	31.22	92.00	61.61
SYNTH-026	89.83 $\pm$ 1.89	89.17 $\pm$ 1.26	89.50 $\pm$ 1.73	90.00 $\pm$ 0.87	33.55	90.97	62.26
SYNTH-027	97.50 $\pm$ 2.65	97.50 $\pm$ 2.18	97.83 $\pm$ 2.25	97.17 $\pm$ 1.53	33.07	97.67	65.37
SYNTH-028	96.83 $\pm$ 0.76	97.33 $\pm$ 0.29	97.50 $\pm$ 0.87	96.83 $\pm$ 1.61	35.10	96.10	65.60
SYNTH-029	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	31.58	100.00	65.79
SYNTH-030	99.33 $\pm$ 0.76	99.17 $\pm$ 0.58	99.50 $\pm$ 0.00	99.17 $\pm$ 0.58	33.82	99.60	66.71
SYNTH-031	35.17 $\pm$ 1.76	35.83 $\pm$ 2.02	32.83 $\pm$ 3.33	37.67 $\pm$ 0.76	40.82	33.98	37.40
SYNTH-032	32.33 $\pm$ 1.89	32.83 $\pm$ 3.88	33.83 $\pm$ 3.21	36.00 $\pm$ 2.60	35.47	35.27	35.37
SYNTH-033	35.67 $\pm$ 1.15	38.83 $\pm$ 2.02	36.83 $\pm$ 1.53	35.67 $\pm$ 2.08	36.40	37.07	36.73
SYNTH-034	37.33 $\pm$ 4.19	34.17 $\pm$ 2.75	37.17 $\pm$ 2.25	35.17 $\pm$ 2.75	38.17	33.70	35.93
SYNTH-035	38.00 $\pm$ 1.80	35.17 $\pm$ 5.80	38.33 $\pm$ 3.21	38.00 $\pm$ 0.50	35.77	41.72	38.74
SYNTH-036	45.33 $\pm$ 2.25	45.67 $\pm$ 2.02	44.00 $\pm$ 4.00	42.67 $\pm$ 7.25	37.38	44.12	40.75
SYNTH-037	45.33 $\pm$ 2.84	48.50 $\pm$ 3.00	46.00 $\pm$ 0.50	46.50 $\pm$ 6.76	41.77	50.12	45.94
SYNTH-038	50.83 $\pm$ 2.93	46.83 $\pm$ 1.76	48.50 $\pm$ 3.04	47.83 $\pm$ 4.04	39.45	49.85	44.65
SYNTH-039	53.33 $\pm$ 1.89	51.00 $\pm$ 3.00	54.00 $\pm$ 2.78	49.50 $\pm$ 2.29	36.45	52.08	44.27
SYNTH-040	48.50 $\pm$ 4.27	47.67 $\pm$ 2.25	48.67 $\pm$ 4.31	51.17 $\pm$ 6.11	37.47	53.75	45.61
SYNTH-041	55.67 $\pm$ 3.40	57.17 $\pm$ 4.25	60.17 $\pm$ 2.75	56.83 $\pm$ 3.25	36.72	57.20	46.96
SYNTH-042	57.17 $\pm$ 1.76	58.00 $\pm$ 5.22	56.67 $\pm$ 0.76	56.17 $\pm$ 2.25	40.32	58.27	49.29
SYNTH-043	64.33 $\pm$ 2.25	62.33 $\pm$ 6.90	60.50 $\pm$ 4.09	62.17 $\pm$ 7.18	38.03	64.12	51.07
SYNTH-044	57.17 $\pm$ 1.53	60.83 $\pm$ 3.51	60.50 $\pm$ 0.87	59.00 $\pm$ 3.77	35.92	61.83	48.88
SYNTH-045	68.17 $\pm$ 2.02	66.00 $\pm$ 1.32	66.67 $\pm$ 1.61	67.00 $\pm$ 1.73	35.95	66.67	51.31
SYNTH-046	73.83 $\pm$ 2.84	70.67 $\pm$ 2.75	74.17 $\pm$ 3.25	69.00 $\pm$ 2.18	38.10	71.07	54.58

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-047	75.00 $\pm$ 3.97	76.50 $\pm$ 3.28	76.83 $\pm$ 1.76	73.83 $\pm$ 1.26	38.58	72.27	55.42
SYNTH-048	73.33 $\pm$ 0.29	73.17 $\pm$ 2.52	73.17 $\pm$ 3.33	72.33 $\pm$ 1.61	37.75	71.97	54.86
SYNTH-049	76.00 $\pm$ 2.00	75.67 $\pm$ 0.76	74.00 $\pm$ 3.04	76.83 $\pm$ 0.29	38.85	74.92	56.88
SYNTH-050	78.17 $\pm$ 4.65	77.17 $\pm$ 2.36	75.50 $\pm$ 2.78	76.50 $\pm$ 2.65	37.48	77.88	57.68
SYNTH-051	83.17 $\pm$ 2.02	82.17 $\pm$ 3.06	83.83 $\pm$ 4.07	82.17 $\pm$ 3.88	35.33	82.25	58.79
SYNTH-052	82.67 $\pm$ 4.62	82.50 $\pm$ 6.24	82.67 $\pm$ 5.58	82.50 $\pm$ 5.68	36.52	80.90	58.71
SYNTH-053	83.67 $\pm$ 1.04	85.67 $\pm$ 2.02	84.67 $\pm$ 2.47	84.33 $\pm$ 2.47	34.40	85.27	59.83
SYNTH-054	85.00 $\pm$ 4.00	87.00 $\pm$ 4.09	85.33 $\pm$ 3.88	86.00 $\pm$ 4.09	38.50	85.42	61.96
SYNTH-055	90.67 $\pm$ 0.58	91.33 $\pm$ 2.25	92.67 $\pm$ 1.04	91.67 $\pm$ 2.52	38.13	91.45	64.79
SYNTH-056	91.83 $\pm$ 2.36	91.17 $\pm$ 1.76	91.83 $\pm$ 1.89	92.33 $\pm$ 1.53	35.10	91.90	63.50
SYNTH-057	90.17 $\pm$ 2.47	90.50 $\pm$ 2.29	90.83 $\pm$ 3.18	90.67 $\pm$ 2.47	39.83	92.22	66.02
SYNTH-058	92.83 $\pm$ 1.61	92.50 $\pm$ 2.18	93.17 $\pm$ 1.04	92.50 $\pm$ 1.80	37.08	94.52	65.80
SYNTH-059	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	40.15	99.90	70.03
SYNTH-060	96.50 $\pm$ 1.32	96.50 $\pm$ 1.32	96.50 $\pm$ 1.32	96.83 $\pm$ 1.26	37.37	97.65	67.51
SYNTH-061	32.50 $\pm$ 3.28	32.50 $\pm$ 3.04	34.00 $\pm$ 1.73	29.00 $\pm$ 2.18	43.05	30.68	36.87
SYNTH-062	35.33 $\pm$ 2.02	37.33 $\pm$ 1.26	37.83 $\pm$ 1.44	31.00 $\pm$ 1.50	44.90	34.13	39.52
SYNTH-063	38.67 $\pm$ 5.03	37.17 $\pm$ 3.33	34.50 $\pm$ 5.77	34.83 $\pm$ 5.20	43.85	36.13	39.99
SYNTH-064	41.83 $\pm$ 1.26	39.83 $\pm$ 3.82	39.00 $\pm$ 4.92	37.33 $\pm$ 2.89	43.43	36.60	40.02
SYNTH-065	43.83 $\pm$ 1.76	49.00 $\pm$ 2.00	43.67 $\pm$ 1.53	43.67 $\pm$ 2.52	40.60	44.80	42.70
SYNTH-066	41.17 $\pm$ 2.36	42.67 $\pm$ 4.86	43.83 $\pm$ 2.08	41.17 $\pm$ 2.02	39.42	40.65	40.03
SYNTH-067	49.00 $\pm$ 8.00	50.33 $\pm$ 2.25	52.50 $\pm$ 2.29	47.67 $\pm$ 4.16	45.40	50.62	48.01
SYNTH-068	49.67 $\pm$ 2.02	47.50 $\pm$ 2.65	50.17 $\pm$ 2.36	49.67 $\pm$ 2.75	40.22	49.35	44.78
SYNTH-069	48.50 $\pm$ 1.00	50.50 $\pm$ 4.58	45.17 $\pm$ 3.33	50.00 $\pm$ 3.28	42.55	52.87	47.71
SYNTH-070	52.33 $\pm$ 3.21	51.50 $\pm$ 1.80	54.17 $\pm$ 1.53	54.67 $\pm$ 5.01	45.52	51.92	48.72
SYNTH-071	59.33 $\pm$ 1.44	60.67 $\pm$ 0.76	56.17 $\pm$ 2.25	58.33 $\pm$ 5.80	42.50	57.65	50.07
SYNTH-072	58.67 $\pm$ 1.26	56.17 $\pm$ 1.26	57.17 $\pm$ 3.33	55.67 $\pm$ 2.08	46.40	56.58	51.49
SYNTH-073	61.83 $\pm$ 2.47	61.00 $\pm$ 2.29	63.67 $\pm$ 2.36	63.17 $\pm$ 3.40	42.43	64.28	53.36
SYNTH-074	59.83 $\pm$ 3.21	61.33 $\pm$ 2.02	61.17 $\pm$ 3.88	62.00 $\pm$ 1.73	43.18	60.08	51.63
SYNTH-075	65.50 $\pm$ 3.91	66.67 $\pm$ 3.82	70.50 $\pm$ 3.12	65.83 $\pm$ 2.57	45.20	69.27	57.23
SYNTH-076	67.50 $\pm$ 4.44	68.83 $\pm$ 4.86	69.17 $\pm$ 0.76	68.50 $\pm$ 1.80	42.30	69.18	55.74
SYNTH-077	67.83 $\pm$ 2.89	70.17 $\pm$ 6.17	67.33 $\pm$ 1.61	68.50 $\pm$ 6.06	42.02	68.83	55.42
SYNTH-078	73.17 $\pm$ 3.21	72.00 $\pm$ 4.92	73.17 $\pm$ 4.48	73.83 $\pm$ 5.01	40.27	72.52	56.39
SYNTH-079	77.00 $\pm$ 4.92	78.67 $\pm$ 1.44	78.00 $\pm$ 2.65	77.50 $\pm$ 4.27	43.20	76.72	59.96
SYNTH-080	77.83 $\pm$ 0.76	77.67 $\pm$ 0.29	77.83 $\pm$ 2.47	78.00 $\pm$ 0.87	44.28	77.03	60.66
SYNTH-081	79.50 $\pm$ 2.60	79.83 $\pm$ 4.04	78.17 $\pm$ 1.26	79.17 $\pm$ 1.89	43.83	81.47	62.65
SYNTH-082	83.50 $\pm$ 4.09	82.50 $\pm$ 1.73	84.67 $\pm$ 2.75	83.17 $\pm$ 2.75	40.70	84.78	62.74
SYNTH-083	88.50 $\pm$ 3.91	88.00 $\pm$ 4.09	87.83 $\pm$ 2.75	88.00 $\pm$ 3.28	41.83	88.22	65.03
SYNTH-084	83.17 $\pm$ 4.07	84.50 $\pm$ 2.50	83.00 $\pm$ 3.04	83.33 $\pm$ 2.47	42.70	85.00	63.85
SYNTH-085	92.83 $\pm$ 1.53	92.67 $\pm$ 1.61	92.67 $\pm$ 1.04	92.33 $\pm$ 1.04	45.97	92.37	69.17
SYNTH-086	92.83 $\pm$ 1.15	93.17 $\pm$ 2.02	92.33 $\pm$ 1.04	92.67 $\pm$ 1.61	44.93	91.87	68.40
SYNTH-087	93.00 $\pm$ 1.32	93.83 $\pm$ 1.44	93.33 $\pm$ 0.76	92.67 $\pm$ 1.04	42.47	94.12	68.29
SYNTH-088	95.00 $\pm$ 1.32	94.33 $\pm$ 2.08	93.00 $\pm$ 3.61	93.33 $\pm$ 3.06	43.92	95.07	69.49
SYNTH-089	98.17 $\pm$ 0.76	98.33 $\pm$ 0.58	98.33 $\pm$ 0.58	98.33 $\pm$ 0.29	39.72	98.78	69.25
SYNTH-090	99.83 $\pm$ 0.29	99.67 $\pm$ 0.29	99.67 $\pm$ 0.29	99.67 $\pm$ 0.29	45.47	99.70	72.58
SYNTH-091	43.33 $\pm$ 5.11	40.33 $\pm$ 4.86	36.50 $\pm$ 3.12	34.00 $\pm$ 4.33	47.97	36.72	42.34
SYNTH-092	38.33 $\pm$ 6.33	37.00 $\pm$ 3.28	39.00 $\pm$ 3.50	32.17 $\pm$ 3.25	47.33	34.67	41.00
SYNTH-093	40.67 $\pm$ 2.52	40.83 $\pm$ 3.33	40.17 $\pm$ 6.25	37.33 $\pm$ 2.84	47.73	39.03	43.38
SYNTH-094	41.17 $\pm$ 1.44	40.17 $\pm$ 2.25	38.33 $\pm$ 2.47	34.83 $\pm$ 1.53	45.73	37.28	41.51
SYNTH-095	48.83 $\pm$ 8.50	42.33 $\pm$ 2.08	40.33 $\pm$ 3.18	39.50 $\pm$ 4.82	46.57	41.17	43.87
SYNTH-096	46.17 $\pm$ 1.04	45.00 $\pm$ 1.80	42.67 $\pm$ 1.76	43.33 $\pm$ 2.57	46.15	40.20	43.17
SYNTH-097	50.67 $\pm$ 1.26	49.33 $\pm$ 3.33	45.50 $\pm$ 0.50	46.67 $\pm$ 1.26	48.45	46.43	47.44
SYNTH-098	52.33 $\pm$ 2.75	50.50 $\pm$ 3.04	45.67 $\pm$ 2.02	43.00 $\pm$ 2.29	49.67	45.82	47.74
SYNTH-099	55.83 $\pm$ 1.76	52.83 $\pm$ 4.31	50.50 $\pm$ 1.32	50.33 $\pm$ 4.65	45.18	51.17	48.17
SYNTH-100	54.67 $\pm$ 3.33	51.17 $\pm$ 0.76	51.17 $\pm$ 1.04	52.17 $\pm$ 1.04	50.27	52.33	51.30
SYNTH-101	56.17 $\pm$ 5.75	58.17 $\pm$ 2.02	57.83 $\pm$ 1.04	53.83 $\pm$ 4.51	48.40	57.05	52.72
SYNTH-102	62.33 $\pm$ 0.58	61.33 $\pm$ 2.84	57.33 $\pm$ 0.76	56.00 $\pm$ 0.50	48.50	56.45	52.48
SYNTH-103	59.83 $\pm$ 3.33	62.00 $\pm$ 2.60	61.50 $\pm$ 5.27	62.67 $\pm$ 5.80	46.20	59.15	52.67
SYNTH-104	63.17 $\pm$ 2.08	64.67 $\pm$ 2.02	64.50 $\pm$ 1.50	63.33 $\pm$ 4.01	41.33	62.65	51.99

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-105	62.00 $\pm$ 2.65	63.17 $\pm$ 2.47	61.33 $\pm$ 2.84	61.17 $\pm$ 2.75	44.00	62.95	53.47
SYNTH-106	71.17 $\pm$ 0.76	70.33 $\pm$ 0.76	69.50 $\pm$ 1.50	70.50 $\pm$ 1.00	48.73	67.87	58.30
SYNTH-107	74.33 $\pm$ 2.47	70.50 $\pm$ 4.27	71.67 $\pm$ 3.51	72.50 $\pm$ 4.58	47.30	72.72	60.01
SYNTH-108	73.83 $\pm$ 1.53	73.00 $\pm$ 1.32	72.67 $\pm$ 0.76	73.00 $\pm$ 2.29	48.27	74.47	61.37
SYNTH-109	76.83 $\pm$ 1.89	77.67 $\pm$ 1.76	77.67 $\pm$ 3.69	77.00 $\pm$ 0.87	46.02	76.40	61.21
SYNTH-110	75.67 $\pm$ 2.47	73.50 $\pm$ 1.32	73.00 $\pm$ 0.87	75.17 $\pm$ 1.53	48.40	75.28	61.84
SYNTH-111	79.67 $\pm$ 2.31	78.17 $\pm$ 1.76	79.67 $\pm$ 0.29	78.00 $\pm$ 2.29	44.00	81.45	62.73
SYNTH-112	77.83 $\pm$ 3.21	79.33 $\pm$ 3.82	77.67 $\pm$ 1.26	76.83 $\pm$ 1.89	45.40	78.60	62.00
SYNTH-113	83.17 $\pm$ 1.61	84.67 $\pm$ 0.29	84.17 $\pm$ 2.08	83.00 $\pm$ 2.18	48.02	83.82	65.92
SYNTH-114	86.83 $\pm$ 1.44	86.50 $\pm$ 1.32	86.67 $\pm$ 1.53	86.67 $\pm$ 0.58	47.58	88.47	68.03
SYNTH-115	90.00 $\pm$ 2.00	90.50 $\pm$ 2.29	90.50 $\pm$ 3.00	90.50 $\pm$ 1.32	49.10	91.30	70.20
SYNTH-116	87.50 $\pm$ 4.58	87.00 $\pm$ 3.12	85.50 $\pm$ 3.12	88.33 $\pm$ 2.25	48.57	88.25	68.41
SYNTH-117	96.00 $\pm$ 1.32	96.00 $\pm$ 1.73	96.00 $\pm$ 1.32	96.50 $\pm$ 1.50	43.88	96.32	70.10
SYNTH-118	94.83 $\pm$ 1.44	94.17 $\pm$ 1.76	94.17 $\pm$ 1.76	93.00 $\pm$ 1.32	43.48	95.07	69.27
SYNTH-119	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	47.12	100.00	73.56
SYNTH-120	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	47.13	100.00	73.57
SYNTH-121	39.33 $\pm$ 6.51	37.00 $\pm$ 2.50	36.00 $\pm$ 2.65	30.50 $\pm$ 2.00	51.67	35.07	43.37
SYNTH-122	47.33 $\pm$ 2.36	43.00 $\pm$ 3.12	38.17 $\pm$ 2.52	37.33 $\pm$ 2.02	54.03	32.67	43.35
SYNTH-123	46.17 $\pm$ 5.20	39.33 $\pm$ 1.26	37.83 $\pm$ 1.76	36.33 $\pm$ 2.36	53.40	37.12	45.26
SYNTH-124	52.50 $\pm$ 3.00	44.67 $\pm$ 2.36	43.00 $\pm$ 5.00	41.33 $\pm$ 5.48	53.80	40.37	47.08
SYNTH-125	52.67 $\pm$ 2.36	49.17 $\pm$ 4.51	44.33 $\pm$ 4.65	41.83 $\pm$ 2.52	53.88	45.72	49.80
SYNTH-126	49.83 $\pm$ 8.50	48.50 $\pm$ 1.80	44.67 $\pm$ 4.25	44.00 $\pm$ 4.92	53.58	42.80	48.19
SYNTH-127	49.67 $\pm$ 4.07	47.17 $\pm$ 3.62	48.17 $\pm$ 3.51	43.17 $\pm$ 0.29	51.42	43.93	47.67
SYNTH-128	53.33 $\pm$ 4.04	47.67 $\pm$ 2.52	50.33 $\pm$ 1.04	50.00 $\pm$ 1.73	49.87	46.00	47.93
SYNTH-129	55.83 $\pm$ 5.84	52.83 $\pm$ 3.75	51.17 $\pm$ 8.46	50.83 $\pm$ 2.08	51.98	51.52	51.75
SYNTH-130	58.00 $\pm$ 0.87	51.83 $\pm$ 2.84	51.00 $\pm$ 1.32	52.00 $\pm$ 4.44	50.80	52.52	51.66
SYNTH-131	62.33 $\pm$ 5.25	59.83 $\pm$ 2.75	57.17 $\pm$ 1.89	57.33 $\pm$ 1.76	49.42	59.13	54.28
SYNTH-132	63.00 $\pm$ 2.78	61.00 $\pm$ 1.00	64.00 $\pm$ 3.04	59.33 $\pm$ 1.04	52.93	60.87	56.90
SYNTH-133	65.50 $\pm$ 3.28	63.67 $\pm$ 2.31	62.83 $\pm$ 1.53	64.00 $\pm$ 3.50	50.18	63.10	56.64
SYNTH-134	68.50 $\pm$ 6.50	63.00 $\pm$ 1.50	60.67 $\pm$ 4.80	61.50 $\pm$ 4.33	54.68	60.17	57.42
SYNTH-135	73.83 $\pm$ 1.76	67.83 $\pm$ 1.76	66.33 $\pm$ 4.80	67.33 $\pm$ 0.58	51.95	65.17	58.56
SYNTH-136	70.33 $\pm$ 2.93	69.67 $\pm$ 3.25	70.67 $\pm$ 3.18	69.17 $\pm$ 2.52	52.93	68.23	60.58
SYNTH-137	74.33 $\pm$ 2.25	70.83 $\pm$ 4.04	73.00 $\pm$ 5.07	73.00 $\pm$ 3.46	47.70	71.97	59.83
SYNTH-138	70.67 $\pm$ 3.25	72.67 $\pm$ 3.40	71.50 $\pm$ 2.65	73.00 $\pm$ 1.50	54.70	71.62	63.16
SYNTH-139	74.50 $\pm$ 1.80	76.00 $\pm$ 3.04	73.50 $\pm$ 5.57	73.67 $\pm$ 3.01	50.82	74.77	62.79
SYNTH-140	76.00 $\pm$ 1.80	73.67 $\pm$ 3.75	73.67 $\pm$ 4.37	73.50 $\pm$ 4.44	50.15	73.18	61.67
SYNTH-141	79.50 $\pm$ 2.00	80.67 $\pm$ 1.26	77.00 $\pm$ 1.32	79.17 $\pm$ 0.58	52.95	79.58	66.27
SYNTH-142	84.17 $\pm$ 3.18	82.83 $\pm$ 2.47	83.67 $\pm$ 2.75	85.00 $\pm$ 1.32	51.32	80.27	65.79
SYNTH-143	82.50 $\pm$ 1.00	82.50 $\pm$ 2.18	83.50 $\pm$ 1.80	82.17 $\pm$ 3.33	52.28	83.08	67.68
SYNTH-144	87.50 $\pm$ 3.00	86.83 $\pm$ 0.76	85.83 $\pm$ 4.01	86.17 $\pm$ 3.75	50.48	85.48	67.98
SYNTH-145	90.50 $\pm$ 2.29	91.17 $\pm$ 2.36	91.67 $\pm$ 2.84	90.67 $\pm$ 2.36	54.73	90.05	72.39
SYNTH-146	92.00 $\pm$ 2.65	92.33 $\pm$ 1.04	92.17 $\pm$ 2.25	91.83 $\pm$ 1.04	51.00	90.75	70.88
SYNTH-147	94.67 $\pm$ 0.76	95.83 $\pm$ 1.26	94.67 $\pm$ 0.76	95.17 $\pm$ 0.29	51.13	94.88	73.01
SYNTH-148	95.17 $\pm$ 1.61	94.50 $\pm$ 0.50	95.67 $\pm$ 0.29	94.50 $\pm$ 0.50	55.68	94.77	75.22
SYNTH-149	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	51.65	100.00	75.83
SYNTH-150	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	52.88	100.00	76.44
SYNTH-151	58.33 $\pm$ 3.51	45.67 $\pm$ 4.16	40.33 $\pm$ 4.25	34.50 $\pm$ 4.27	61.92	36.40	49.16
SYNTH-152	48.67 $\pm$ 2.36	41.00 $\pm$ 6.06	38.67 $\pm$ 2.02	35.50 $\pm$ 4.09	57.85	32.68	45.27
SYNTH-153	53.00 $\pm$ 4.44	50.50 $\pm$ 3.28	44.00 $\pm$ 1.80	38.83 $\pm$ 5.53	56.78	39.07	47.93
SYNTH-154	51.67 $\pm$ 4.73	40.50 $\pm$ 5.89	41.00 $\pm$ 5.41	36.00 $\pm$ 2.60	56.72	32.83	44.77
SYNTH-155	50.50 $\pm$ 1.00	42.83 $\pm$ 4.19	43.17 $\pm$ 1.89	39.00 $\pm$ 2.60	55.32	42.68	49.00
SYNTH-156	53.00 $\pm$ 4.09	47.17 $\pm$ 1.76	44.17 $\pm$ 4.07	45.00 $\pm$ 2.60	58.68	43.05	50.87
SYNTH-157	59.83 $\pm$ 1.44	54.00 $\pm$ 3.28	49.33 $\pm$ 1.61	49.00 $\pm$ 3.61	54.72	49.40	52.06
SYNTH-158	59.33 $\pm$ 1.53	51.50 $\pm$ 1.73	49.50 $\pm$ 2.65	45.83 $\pm$ 3.69	59.07	47.58	53.33
SYNTH-159	61.00 $\pm$ 2.65	50.50 $\pm$ 3.91	50.67 $\pm$ 4.07	51.50 $\pm$ 1.32	55.63	51.13	53.38
SYNTH-160	60.50 $\pm$ 2.78	57.00 $\pm$ 6.26	56.33 $\pm$ 2.47	55.83 $\pm$ 3.62	57.15	51.75	54.45
SYNTH-161	65.17 $\pm$ 1.89	56.83 $\pm$ 2.25	57.17 $\pm$ 5.11	55.83 $\pm$ 2.36	58.63	58.25	58.44
SYNTH-162	67.33 $\pm$ 6.75	59.00 $\pm$ 3.61	58.67 $\pm$ 3.55	56.83 $\pm$ 4.91	58.20	58.68	58.44

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-163	67.00 $\pm$ 1.32	64.00 $\pm$ 2.00	64.00 $\pm$ 3.91	58.67 $\pm$ 2.25	55.80	59.02	57.41
SYNTH-164	70.33 $\pm$ 0.58	63.83 $\pm$ 0.76	61.33 $\pm$ 2.36	62.17 $\pm$ 4.16	57.70	59.55	58.62
SYNTH-165	70.17 $\pm$ 3.55	62.33 $\pm$ 4.65	66.83 $\pm$ 2.57	61.67 $\pm$ 4.51	56.47	65.83	61.15
SYNTH-166	71.00 $\pm$ 2.65	65.17 $\pm$ 2.47	63.67 $\pm$ 0.58	61.33 $\pm$ 5.03	57.03	63.33	60.18
SYNTH-167	74.83 $\pm$ 1.26	74.00 $\pm$ 0.00	71.67 $\pm$ 1.61	69.50 $\pm$ 1.32	52.43	73.27	62.85
SYNTH-168	75.33 $\pm$ 2.75	70.67 $\pm$ 2.31	71.00 $\pm$ 0.87	69.00 $\pm$ 3.91	55.62	69.72	62.67
SYNTH-169	77.33 $\pm$ 1.76	74.83 $\pm$ 2.36	74.67 $\pm$ 2.47	75.00 $\pm$ 0.50	60.08	73.72	66.90
SYNTH-170	80.67 $\pm$ 1.53	76.17 $\pm$ 3.33	77.17 $\pm$ 1.89	76.83 $\pm$ 2.36	58.28	76.42	67.35
SYNTH-171	79.50 $\pm$ 1.50	81.33 $\pm$ 4.31	80.83 $\pm$ 0.58	79.67 $\pm$ 0.29	58.73	80.28	69.51
SYNTH-172	83.00 $\pm$ 2.29	80.50 $\pm$ 0.00	80.50 $\pm$ 0.87	79.83 $\pm$ 2.47	56.52	80.75	68.63
SYNTH-173	86.17 $\pm$ 1.26	86.50 $\pm$ 1.32	86.67 $\pm$ 1.26	86.00 $\pm$ 1.80	56.03	85.97	71.00
SYNTH-174	84.67 $\pm$ 4.04	84.50 $\pm$ 2.78	85.17 $\pm$ 3.55	84.17 $\pm$ 2.25	57.60	85.78	71.69
SYNTH-175	90.33 $\pm$ 1.26	90.83 $\pm$ 0.76	89.17 $\pm$ 1.15	91.17 $\pm$ 1.44	59.18	88.72	73.95
SYNTH-176	88.67 $\pm$ 2.75	90.00 $\pm$ 2.00	89.67 $\pm$ 2.25	89.50 $\pm$ 1.73	58.20	90.07	74.13
SYNTH-177	97.00 $\pm$ 1.32	96.17 $\pm$ 0.76	96.17 $\pm$ 2.02	96.33 $\pm$ 0.76	58.10	96.43	77.27
SYNTH-178	95.50 $\pm$ 2.00	95.33 $\pm$ 1.61	95.50 $\pm$ 1.50	95.33 $\pm$ 1.26	57.38	95.97	76.67
SYNTH-179	99.50 $\pm$ 0.50	99.50 $\pm$ 0.00	99.50 $\pm$ 0.00	99.50 $\pm$ 0.00	54.83	99.55	77.19
SYNTH-180	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	57.85	100.00	78.92
SYNTH-181	55.67 $\pm$ 2.08	48.17 $\pm$ 1.89	40.83 $\pm$ 5.01	35.00 $\pm$ 3.61	62.60	33.33	47.97
SYNTH-182	47.67 $\pm$ 3.82	45.00 $\pm$ 4.44	33.50 $\pm$ 5.41	34.17 $\pm$ 3.79	63.27	36.08	49.67
SYNTH-183	48.50 $\pm$ 6.26	44.67 $\pm$ 2.08	40.50 $\pm$ 1.80	37.50 $\pm$ 1.00	63.02	41.07	52.04
SYNTH-184	52.67 $\pm$ 0.76	46.00 $\pm$ 1.73	43.17 $\pm$ 1.53	38.00 $\pm$ 2.29	60.38	37.05	48.72
SYNTH-185	62.50 $\pm$ 4.50	54.00 $\pm$ 1.80	48.00 $\pm$ 3.46	43.33 $\pm$ 1.04	62.95	44.43	53.69
SYNTH-186	54.33 $\pm$ 4.54	45.17 $\pm$ 1.76	43.00 $\pm$ 1.00	41.67 $\pm$ 3.21	62.37	42.13	52.25
SYNTH-187	60.00 $\pm$ 1.73	52.50 $\pm$ 2.29	50.00 $\pm$ 3.77	41.17 $\pm$ 2.52	65.02	46.23	55.62
SYNTH-188	66.83 $\pm$ 1.04	54.17 $\pm$ 4.16	48.83 $\pm$ 0.76	46.83 $\pm$ 2.47	63.28	48.08	55.68
SYNTH-189	66.83 $\pm$ 5.11	59.83 $\pm$ 4.54	58.67 $\pm$ 3.21	55.00 $\pm$ 4.58	61.25	52.65	56.95
SYNTH-190	67.00 $\pm$ 4.09	61.00 $\pm$ 2.18	55.67 $\pm$ 2.36	56.17 $\pm$ 2.47	61.77	54.72	58.24
SYNTH-191	71.33 $\pm$ 4.07	65.33 $\pm$ 2.84	61.00 $\pm$ 2.18	58.33 $\pm$ 2.93	64.17	59.92	62.04
SYNTH-192	67.33 $\pm$ 0.29	60.50 $\pm$ 0.50	56.67 $\pm$ 2.02	56.67 $\pm$ 4.04	62.77	54.92	58.84
SYNTH-193	72.67 $\pm$ 3.62	68.50 $\pm$ 5.27	67.00 $\pm$ 0.87	66.50 $\pm$ 1.32	60.88	60.55	60.72
SYNTH-194	71.67 $\pm$ 1.15	63.67 $\pm$ 1.53	64.67 $\pm$ 1.76	58.67 $\pm$ 0.58	61.57	60.92	61.24
SYNTH-195	77.17 $\pm$ 5.11	72.83 $\pm$ 1.15	71.33 $\pm$ 4.75	66.83 $\pm$ 5.80	60.63	67.85	64.24
SYNTH-196	78.67 $\pm$ 0.58	75.50 $\pm$ 5.07	74.17 $\pm$ 1.89	73.67 $\pm$ 3.82	62.95	70.63	66.79
SYNTH-197	80.00 $\pm$ 1.00	73.50 $\pm$ 3.04	73.17 $\pm$ 0.76	71.50 $\pm$ 1.32	61.80	73.23	67.52
SYNTH-198	77.00 $\pm$ 3.61	73.00 $\pm$ 5.07	72.17 $\pm$ 4.19	71.83 $\pm$ 4.93	63.82	71.23	67.53
SYNTH-199	81.33 $\pm$ 2.89	78.17 $\pm$ 2.25	78.50 $\pm$ 1.50	77.50 $\pm$ 1.80	61.02	78.00	69.51
SYNTH-200	82.17 $\pm$ 1.26	78.67 $\pm$ 2.25	77.00 $\pm$ 1.00	75.67 $\pm$ 0.58	64.18	76.40	70.29
SYNTH-201	79.67 $\pm$ 1.61	86.83 $\pm$ 0.76	86.00 $\pm$ 1.32	84.67 $\pm$ 1.76	63.72	82.98	73.35
SYNTH-202	79.17 $\pm$ 2.75	79.67 $\pm$ 2.25	79.17 $\pm$ 2.75	79.33 $\pm$ 1.26	59.78	78.08	68.93
SYNTH-203	88.00 $\pm$ 2.78	85.50 $\pm$ 2.00	87.50 $\pm$ 2.78	87.17 $\pm$ 2.57	63.85	86.30	75.08
SYNTH-204	85.67 $\pm$ 1.61	84.67 $\pm$ 0.76	86.00 $\pm$ 1.32	86.50 $\pm$ 3.46	60.75	86.57	73.66
SYNTH-205	91.67 $\pm$ 1.61	90.83 $\pm$ 1.26	90.17 $\pm$ 0.29	90.00 $\pm$ 1.00	60.50	90.10	75.30
SYNTH-206	90.67 $\pm$ 1.61	90.33 $\pm$ 1.61	89.67 $\pm$ 2.36	90.83 $\pm$ 0.58	64.47	90.57	77.52
SYNTH-207	94.00 $\pm$ 0.50	94.67 $\pm$ 1.53	94.50 $\pm$ 0.50	94.33 $\pm$ 1.26	60.58	94.58	77.58
SYNTH-208	94.50 $\pm$ 1.32	93.17 $\pm$ 1.26	93.50 $\pm$ 1.32	93.67 $\pm$ 1.26	59.27	94.18	76.72
SYNTH-209	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	61.20	100.00	80.60
SYNTH-210	99.17 $\pm$ 0.76	99.00 $\pm$ 0.50	99.17 $\pm$ 0.76	99.00 $\pm$ 0.50	60.07	99.07	79.57
SYNTH-211	54.83 $\pm$ 4.54	47.67 $\pm$ 5.13	40.83 $\pm$ 5.35	33.50 $\pm$ 5.07	64.33	32.43	48.38
SYNTH-212	60.00 $\pm$ 5.20	54.67 $\pm$ 4.65	45.17 $\pm$ 1.26	34.83 $\pm$ 5.01	67.60	32.43	50.02
SYNTH-213	62.33 $\pm$ 1.26	55.17 $\pm$ 0.29	46.83 $\pm$ 4.01	42.17 $\pm$ 5.80	68.07	36.02	52.04
SYNTH-214	66.17 $\pm$ 1.76	60.17 $\pm$ 1.53	48.83 $\pm$ 2.02	39.50 $\pm$ 3.50	67.70	40.77	54.23
SYNTH-215	62.17 $\pm$ 1.61	57.83 $\pm$ 3.33	46.17 $\pm$ 1.15	41.17 $\pm$ 2.52	64.22	41.07	52.64
SYNTH-216	62.00 $\pm$ 1.32	55.50 $\pm$ 3.12	47.67 $\pm$ 1.76	41.00 $\pm$ 4.44	67.85	42.35	55.10
SYNTH-217	66.33 $\pm$ 1.26	61.33 $\pm$ 1.53	52.33 $\pm$ 4.37	50.17 $\pm$ 3.75	67.55	49.12	58.33
SYNTH-218	65.17 $\pm$ 0.76	61.00 $\pm$ 1.32	54.67 $\pm$ 3.06	46.33 $\pm$ 4.93	66.85	45.70	56.27
SYNTH-219	68.17 $\pm$ 2.08	60.17 $\pm$ 2.52	54.83 $\pm$ 2.93	52.50 $\pm$ 0.50	65.58	56.45	61.02
SYNTH-220	70.17 $\pm$ 2.75	63.50 $\pm$ 2.29	60.00 $\pm$ 0.87	56.33 $\pm$ 1.26	66.50	56.23	61.37

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-221	73.33 $\pm$ 0.76	65.50 $\pm$ 2.18	64.83 $\pm$ 4.54	59.83 $\pm$ 1.04	66.45	58.63	62.54
SYNTH-222	73.33 $\pm$ 2.08	64.00 $\pm$ 1.32	60.33 $\pm$ 3.75	56.17 $\pm$ 1.76	67.52	56.82	62.17
SYNTH-223	69.33 $\pm$ 6.45	67.00 $\pm$ 8.19	59.17 $\pm$ 5.06	55.83 $\pm$ 3.06	65.70	58.87	62.28
SYNTH-224	73.67 $\pm$ 2.75	66.83 $\pm$ 4.16	62.50 $\pm$ 3.77	62.83 $\pm$ 0.58	65.37	63.82	64.59
SYNTH-225	77.17 $\pm$ 1.44	71.33 $\pm$ 2.47	68.67 $\pm$ 2.02	68.50 $\pm$ 3.04	63.52	66.97	65.24
SYNTH-226	76.17 $\pm$ 3.33	67.17 $\pm$ 2.75	66.67 $\pm$ 1.89	63.00 $\pm$ 0.50	65.17	64.97	65.07
SYNTH-227	79.33 $\pm$ 2.36	72.17 $\pm$ 5.58	72.83 $\pm$ 1.26	72.00 $\pm$ 5.68	68.67	69.52	69.09
SYNTH-228	79.17 $\pm$ 3.01	71.33 $\pm$ 4.75	73.50 $\pm$ 1.50	71.67 $\pm$ 4.25	65.88	71.47	68.67
SYNTH-229	86.33 $\pm$ 2.36	81.33 $\pm$ 3.01	81.50 $\pm$ 2.18	80.67 $\pm$ 3.79	65.43	78.60	72.02
SYNTH-230	80.00 $\pm$ 2.00	73.17 $\pm$ 0.76	72.67 $\pm$ 0.29	72.67 $\pm$ 3.21	65.55	76.05	70.80
SYNTH-231	84.00 $\pm$ 0.87	78.50 $\pm$ 4.33	78.83 $\pm$ 3.01	79.00 $\pm$ 2.65	64.98	79.15	72.07
SYNTH-232	84.83 $\pm$ 1.76	80.33 $\pm$ 1.15	77.33 $\pm$ 3.88	81.33 $\pm$ 1.76	67.10	81.38	74.24
SYNTH-233	86.67 $\pm$ 1.04	83.17 $\pm$ 2.84	83.00 $\pm$ 2.00	84.17 $\pm$ 3.33	68.13	84.50	76.32
SYNTH-234	85.17 $\pm$ 3.06	84.83 $\pm$ 3.69	84.33 $\pm$ 1.26	83.67 $\pm$ 2.02	69.80	85.17	77.48
SYNTH-235	90.00 $\pm$ 2.29	91.17 $\pm$ 2.47	89.50 $\pm$ 1.00	90.00 $\pm$ 1.80	63.98	90.88	77.43
SYNTH-236	90.67 $\pm$ 3.51	89.33 $\pm$ 3.33	89.83 $\pm$ 1.76	89.50 $\pm$ 2.65	66.40	91.03	78.72
SYNTH-237	94.50 $\pm$ 2.00	94.33 $\pm$ 1.76	94.17 $\pm$ 1.53	95.00 $\pm$ 1.80	67.02	95.95	81.48
SYNTH-238	94.83 $\pm$ 0.29	95.33 $\pm$ 0.76	95.17 $\pm$ 0.58	95.33 $\pm$ 1.04	66.45	95.80	81.12
SYNTH-239	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	63.08	99.90	81.49
SYNTH-240	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	64.92	100.00	82.46
SYNTH-241	64.00 $\pm$ 3.91	58.67 $\pm$ 2.36	44.83 $\pm$ 5.06	38.00 $\pm$ 1.73	71.58	34.28	52.93
SYNTH-242	62.50 $\pm$ 2.18	58.17 $\pm$ 1.15	43.00 $\pm$ 3.12	37.50 $\pm$ 2.18	69.90	34.05	51.97
SYNTH-243	68.83 $\pm$ 1.76	61.33 $\pm$ 3.88	45.83 $\pm$ 2.89	32.67 $\pm$ 3.75	71.63	35.03	53.33
SYNTH-244	62.33 $\pm$ 4.54	59.00 $\pm$ 4.44	45.00 $\pm$ 2.60	37.00 $\pm$ 1.32	71.58	39.73	55.66
SYNTH-245	65.50 $\pm$ 2.18	60.83 $\pm$ 1.89	45.50 $\pm$ 3.28	42.33 $\pm$ 1.61	73.93	43.48	58.71
SYNTH-246	65.33 $\pm$ 2.36	62.33 $\pm$ 2.75	48.33 $\pm$ 4.37	42.17 $\pm$ 4.01	69.68	42.68	56.18
SYNTH-247	72.67 $\pm$ 1.44	62.17 $\pm$ 3.21	57.33 $\pm$ 4.75	54.33 $\pm$ 2.36	69.20	48.78	58.99
SYNTH-248	70.67 $\pm$ 2.02	64.33 $\pm$ 1.26	55.00 $\pm$ 5.07	49.83 $\pm$ 4.93	71.12	47.83	59.47
SYNTH-249	76.67 $\pm$ 2.57	71.50 $\pm$ 3.28	58.33 $\pm$ 5.03	53.17 $\pm$ 4.07	74.13	51.83	62.98
SYNTH-250	76.67 $\pm$ 1.15	72.50 $\pm$ 0.00	60.50 $\pm$ 0.50	54.00 $\pm$ 3.04	73.78	53.52	63.65
SYNTH-251	72.67 $\pm$ 5.35	65.33 $\pm$ 1.89	63.67 $\pm$ 2.31	56.67 $\pm$ 3.51	70.73	55.70	63.22
SYNTH-252	76.83 $\pm$ 4.16	68.00 $\pm$ 3.97	61.50 $\pm$ 3.77	58.00 $\pm$ 2.29	70.68	55.95	63.32
SYNTH-253	75.67 $\pm$ 3.51	70.50 $\pm$ 1.80	64.33 $\pm$ 1.76	62.00 $\pm$ 2.29	70.53	60.38	65.46
SYNTH-254	81.17 $\pm$ 1.26	71.50 $\pm$ 1.80	66.67 $\pm$ 0.76	65.50 $\pm$ 1.32	71.65	62.13	66.89
SYNTH-255	83.50 $\pm$ 2.18	77.17 $\pm$ 5.39	69.83 $\pm$ 2.36	70.17 $\pm$ 2.75	71.43	66.03	68.73
SYNTH-256	80.50 $\pm$ 2.00	72.67 $\pm$ 4.04	66.67 $\pm$ 2.08	65.33 $\pm$ 2.84	72.83	65.38	69.11
SYNTH-257	81.33 $\pm$ 1.15	75.50 $\pm$ 1.32	74.00 $\pm$ 5.07	72.33 $\pm$ 2.84	71.67	70.47	71.07
SYNTH-258	77.33 $\pm$ 2.31	71.00 $\pm$ 2.65	69.00 $\pm$ 0.50	66.67 $\pm$ 0.76	70.28	70.98	70.63
SYNTH-259	83.00 $\pm$ 1.80	79.17 $\pm$ 1.53	77.17 $\pm$ 2.75	73.33 $\pm$ 2.75	70.55	74.52	72.53
SYNTH-260	84.50 $\pm$ 1.50	79.33 $\pm$ 1.76	77.67 $\pm$ 2.93	78.17 $\pm$ 2.47	71.05	76.68	73.87
SYNTH-261	85.50 $\pm$ 1.50	81.17 $\pm$ 5.97	78.67 $\pm$ 5.35	81.17 $\pm$ 5.80	69.42	77.18	73.30
SYNTH-262	89.33 $\pm$ 2.02	85.33 $\pm$ 2.47	85.83 $\pm$ 2.02	85.67 $\pm$ 2.02	71.07	82.27	76.67
SYNTH-263	90.50 $\pm$ 3.28	88.17 $\pm$ 2.75	85.83 $\pm$ 3.88	85.00 $\pm$ 1.73	73.10	83.37	78.23
SYNTH-264	90.33 $\pm$ 3.75	91.50 $\pm$ 2.50	88.67 $\pm$ 2.89	89.33 $\pm$ 2.75	71.70	88.22	79.96
SYNTH-265	90.50 $\pm$ 1.32	89.83 $\pm$ 0.29	90.00 $\pm$ 1.73	90.67 $\pm$ 0.58	72.72	89.30	81.01
SYNTH-266	93.33 $\pm$ 1.89	94.17 $\pm$ 1.53	93.33 $\pm$ 1.26	93.50 $\pm$ 1.00	69.25	91.43	80.34
SYNTH-267	96.00 $\pm$ 0.50	96.83 $\pm$ 1.26	96.33 $\pm$ 1.44	96.00 $\pm$ 0.87	70.73	96.35	83.54
SYNTH-268	95.50 $\pm$ 1.32	95.33 $\pm$ 1.26	96.17 $\pm$ 1.15	95.83 $\pm$ 1.26	71.92	95.52	83.72
SYNTH-269	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	72.40	100.00	86.20
SYNTH-270	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	70.98	100.00	85.49
SYNTH-271	66.50 $\pm$ 3.00	59.50 $\pm$ 3.12	45.17 $\pm$ 1.26	31.17 $\pm$ 5.53	73.77	31.13	52.45
SYNTH-272	67.83 $\pm$ 2.52	61.33 $\pm$ 3.69	43.50 $\pm$ 1.50	34.17 $\pm$ 4.54	75.65	31.48	53.57
SYNTH-273	67.67 $\pm$ 2.02	62.33 $\pm$ 2.02	48.17 $\pm$ 2.25	39.67 $\pm$ 4.65	75.07	38.63	56.85
SYNTH-274	66.50 $\pm$ 1.80	65.17 $\pm$ 2.75	49.00 $\pm$ 6.50	43.00 $\pm$ 4.92	75.20	40.50	57.85
SYNTH-275	74.00 $\pm$ 0.50	67.67 $\pm$ 1.61	53.50 $\pm$ 3.91	43.67 $\pm$ 4.65	76.27	42.65	59.46
SYNTH-276	72.00 $\pm$ 3.77	66.50 $\pm$ 2.29	51.33 $\pm$ 1.53	42.83 $\pm$ 4.48	77.30	43.08	60.19
SYNTH-277	74.67 $\pm$ 2.08	69.50 $\pm$ 0.50	53.67 $\pm$ 2.84	49.17 $\pm$ 4.16	74.13	47.90	61.02
SYNTH-278	69.17 $\pm$ 1.53	66.50 $\pm$ 2.60	51.00 $\pm$ 1.00	45.17 $\pm$ 5.48	73.97	49.20	61.58

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-279	77.67 $\pm$ 2.75	70.50 $\pm$ 5.20	57.67 $\pm$ 3.69	56.50 $\pm$ 3.12	74.78	56.55	65.67
SYNTH-280	77.33 $\pm$ 3.01	71.33 $\pm$ 2.31	61.67 $\pm$ 3.69	52.00 $\pm$ 1.50	75.08	51.48	63.28
SYNTH-281	79.67 $\pm$ 3.69	76.67 $\pm$ 1.61	63.33 $\pm$ 3.62	64.00 $\pm$ 0.50	76.20	61.20	68.70
SYNTH-282	81.33 $\pm$ 2.57	74.17 $\pm$ 2.47	67.17 $\pm$ 0.76	61.67 $\pm$ 2.84	79.12	60.78	69.95
SYNTH-283	83.33 $\pm$ 1.15	78.67 $\pm$ 2.36	65.67 $\pm$ 1.61	65.50 $\pm$ 1.50	76.43	63.05	69.74
SYNTH-284	78.00 $\pm$ 1.73	72.83 $\pm$ 2.08	67.17 $\pm$ 3.21	64.00 $\pm$ 3.04	77.15	63.40	70.27
SYNTH-285	81.00 $\pm$ 2.18	74.50 $\pm$ 1.00	65.67 $\pm$ 2.89	67.67 $\pm$ 6.83	75.18	66.20	70.69
SYNTH-286	83.00 $\pm$ 0.50	76.33 $\pm$ 2.02	68.67 $\pm$ 4.16	65.00 $\pm$ 2.78	76.57	66.22	71.39
SYNTH-287	87.50 $\pm$ 0.87	79.50 $\pm$ 3.91	73.33 $\pm$ 0.58	74.83 $\pm$ 2.47	74.47	71.85	73.16
SYNTH-288	83.17 $\pm$ 1.89	75.83 $\pm$ 1.26	70.67 $\pm$ 1.15	67.67 $\pm$ 2.25	77.57	70.68	74.12
SYNTH-289	86.67 $\pm$ 3.21	79.83 $\pm$ 2.47	78.33 $\pm$ 2.36	77.83 $\pm$ 2.02	77.77	77.67	77.72
SYNTH-290	85.50 $\pm$ 1.00	79.17 $\pm$ 1.61	77.17 $\pm$ 2.57	77.67 $\pm$ 2.31	75.63	77.95	76.79
SYNTH-291	87.67 $\pm$ 2.08	83.00 $\pm$ 5.22	81.33 $\pm$ 4.37	81.00 $\pm$ 1.80	75.52	80.70	78.11
SYNTH-292	88.50 $\pm$ 3.28	82.50 $\pm$ 1.00	84.67 $\pm$ 4.16	82.50 $\pm$ 3.91	75.85	80.68	78.27
SYNTH-293	92.33 $\pm$ 2.02	89.00 $\pm$ 1.73	89.83 $\pm$ 1.26	87.17 $\pm$ 1.89	76.47	85.80	81.13
SYNTH-294	90.33 $\pm$ 1.53	88.50 $\pm$ 3.12	87.83 $\pm$ 2.02	87.17 $\pm$ 3.06	75.50	87.85	81.68
SYNTH-295	92.83 $\pm$ 0.29	93.33 $\pm$ 1.15	92.83 $\pm$ 0.29	92.33 $\pm$ 0.76	73.08	91.02	82.05
SYNTH-296	91.67 $\pm$ 1.89	89.50 $\pm$ 1.80	89.33 $\pm$ 1.26	87.17 $\pm$ 1.89	77.42	90.52	83.97
SYNTH-297	94.00 $\pm$ 1.80	94.00 $\pm$ 0.50	94.17 $\pm$ 1.04	93.50 $\pm$ 1.80	75.70	93.10	84.40
SYNTH-298	96.17 $\pm$ 0.58	96.00 $\pm$ 1.32	95.17 $\pm$ 0.76	96.33 $\pm$ 0.29	76.83	95.63	86.23
SYNTH-299	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	77.22	100.00	88.61
SYNTH-300	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	99.83 $\pm$ 0.29	73.07	99.62	86.34
SYNTH-301	73.67 $\pm$ 1.26	71.33 $\pm$ 4.25	47.00 $\pm$ 2.29	38.50 $\pm$ 2.60	81.20	33.47	57.33
SYNTH-302	71.83 $\pm$ 5.84	65.33 $\pm$ 5.39	45.50 $\pm$ 2.50	36.33 $\pm$ 5.20	79.25	31.42	55.33
SYNTH-303	74.00 $\pm$ 1.80	67.33 $\pm$ 2.47	48.67 $\pm$ 6.25	41.83 $\pm$ 1.76	77.57	39.48	58.53
SYNTH-304	75.17 $\pm$ 1.53	71.83 $\pm$ 1.76	53.17 $\pm$ 2.02	42.33 $\pm$ 1.61	82.25	39.05	60.65
SYNTH-305	71.00 $\pm$ 0.50	66.17 $\pm$ 4.48	49.50 $\pm$ 9.53	44.33 $\pm$ 2.93	78.65	43.28	60.97
SYNTH-306	74.83 $\pm$ 3.88	70.33 $\pm$ 2.75	53.17 $\pm$ 2.75	45.83 $\pm$ 2.36	78.60	41.98	60.29
SYNTH-307	75.17 $\pm$ 5.30	69.33 $\pm$ 3.51	55.00 $\pm$ 6.38	52.00 $\pm$ 1.50	79.55	48.35	63.95
SYNTH-308	79.50 $\pm$ 2.78	73.67 $\pm$ 1.53	57.33 $\pm$ 7.37	48.83 $\pm$ 5.62	81.87	45.57	63.72
SYNTH-309	81.17 $\pm$ 4.04	76.50 $\pm$ 2.65	60.83 $\pm$ 7.59	55.50 $\pm$ 1.32	81.48	54.18	67.83
SYNTH-310	77.33 $\pm$ 1.26	69.83 $\pm$ 0.58	57.67 $\pm$ 2.36	52.00 $\pm$ 2.18	79.08	50.63	64.86
SYNTH-311	81.00 $\pm$ 0.50	77.17 $\pm$ 1.15	60.00 $\pm$ 3.00	56.67 $\pm$ 3.79	79.52	54.30	66.91
SYNTH-312	79.17 $\pm$ 4.73	73.00 $\pm$ 7.00	59.83 $\pm$ 3.69	56.83 $\pm$ 4.04	79.55	56.67	68.11
SYNTH-313	83.50 $\pm$ 5.27	80.50 $\pm$ 3.28	69.00 $\pm$ 0.50	69.50 $\pm$ 8.05	80.20	61.40	70.80
SYNTH-314	84.50 $\pm$ 2.00	79.83 $\pm$ 2.25	72.17 $\pm$ 2.75	64.17 $\pm$ 5.86	81.90	61.88	71.89
SYNTH-315	87.67 $\pm$ 2.52	81.33 $\pm$ 3.25	74.67 $\pm$ 4.07	70.50 $\pm$ 3.97	81.17	66.73	73.95
SYNTH-316	89.17 $\pm$ 2.02	82.00 $\pm$ 4.92	71.50 $\pm$ 6.56	65.17 $\pm$ 4.04	81.65	65.68	73.67
SYNTH-317	90.17 $\pm$ 3.33	83.67 $\pm$ 3.25	77.50 $\pm$ 3.61	79.00 $\pm$ 2.18	81.32	74.65	77.98
SYNTH-318	86.17 $\pm$ 1.61	80.33 $\pm$ 4.07	70.83 $\pm$ 3.18	73.17 $\pm$ 2.36	81.23	71.13	76.18
SYNTH-319	87.83 $\pm$ 2.75	83.83 $\pm$ 2.89	80.50 $\pm$ 3.91	80.17 $\pm$ 2.84	80.93	76.22	78.58
SYNTH-320	89.83 $\pm$ 1.61	78.83 $\pm$ 4.31	79.33 $\pm$ 2.93	76.83 $\pm$ 3.33	80.92	75.30	78.11
SYNTH-321	89.17 $\pm$ 1.04	83.17 $\pm$ 4.37	82.00 $\pm$ 0.50	79.33 $\pm$ 2.25	78.43	79.53	78.98
SYNTH-322	92.00 $\pm$ 0.87	84.50 $\pm$ 3.77	82.83 $\pm$ 2.08	83.67 $\pm$ 1.04	81.58	82.05	81.82
SYNTH-323	90.83 $\pm$ 2.36	87.17 $\pm$ 1.76	85.50 $\pm$ 1.32	83.50 $\pm$ 1.50	80.93	84.37	82.65
SYNTH-324	91.17 $\pm$ 3.21	86.83 $\pm$ 3.40	85.33 $\pm$ 1.76	86.00 $\pm$ 1.80	82.52	86.57	84.54
SYNTH-325	93.33 $\pm$ 3.21	90.00 $\pm$ 2.65	90.00 $\pm$ 1.50	88.83 $\pm$ 1.26	80.05	89.17	84.61
SYNTH-326	91.50 $\pm$ 0.50	92.00 $\pm$ 1.50	90.83 $\pm$ 1.53	91.00 $\pm$ 1.32	82.68	91.15	86.92
SYNTH-327	96.50 $\pm$ 0.50	96.00 $\pm$ 0.50	95.83 $\pm$ 0.58	96.17 $\pm$ 0.58	81.15	96.48	88.82
SYNTH-328	97.00 $\pm$ 1.80	96.67 $\pm$ 1.76	96.67 $\pm$ 1.26	96.17 $\pm$ 1.61	79.33	96.68	88.01
SYNTH-329	99.17 $\pm$ 0.76	98.67 $\pm$ 0.58	98.33 $\pm$ 0.58	99.00 $\pm$ 0.50	82.00	99.10	90.55
SYNTH-330	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	80.68	100.00	90.34
SYNTH-331	78.50 $\pm$ 3.12	73.83 $\pm$ 3.62	51.67 $\pm$ 2.02	37.00 $\pm$ 1.73	85.65	33.97	59.81
SYNTH-332	77.33 $\pm$ 2.75	71.33 $\pm$ 2.75	52.50 $\pm$ 5.77	39.17 $\pm$ 2.36	83.98	35.03	59.51
SYNTH-333	81.33 $\pm$ 3.40	76.33 $\pm$ 5.69	55.50 $\pm$ 1.32	42.00 $\pm$ 4.58	87.33	40.23	63.78
SYNTH-334	81.67 $\pm$ 0.76	72.50 $\pm$ 5.63	50.33 $\pm$ 5.97	39.83 $\pm$ 1.04	87.75	41.62	64.68
SYNTH-335	82.67 $\pm$ 4.54	77.83 $\pm$ 4.16	60.00 $\pm$ 4.77	47.83 $\pm$ 3.55	86.87	41.48	64.17
SYNTH-336	84.83 $\pm$ 3.01	81.67 $\pm$ 3.01	57.83 $\pm$ 2.84	48.17 $\pm$ 3.75	87.93	40.38	64.16

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-337	83.83 $\pm$ 2.57	76.83 $\pm$ 2.93	57.83 $\pm$ 3.18	49.83 $\pm$ 4.48	84.88	46.82	65.85
SYNTH-338	84.83 $\pm$ 3.33	79.00 $\pm$ 3.04	58.67 $\pm$ 5.48	49.33 $\pm$ 2.89	86.40	47.93	67.17
SYNTH-339	88.17 $\pm$ 3.55	84.33 $\pm$ 2.25	63.00 $\pm$ 8.41	60.50 $\pm$ 1.00	88.65	50.42	69.53
SYNTH-340	83.67 $\pm$ 2.75	76.17 $\pm$ 0.76	65.00 $\pm$ 4.09	58.50 $\pm$ 2.78	83.10	51.77	67.43
SYNTH-341	85.33 $\pm$ 1.53	81.00 $\pm$ 1.73	68.83 $\pm$ 4.37	64.33 $\pm$ 3.55	84.33	60.40	72.37
SYNTH-342	85.50 $\pm$ 6.08	79.50 $\pm$ 3.91	66.17 $\pm$ 3.88	59.33 $\pm$ 2.36	85.75	56.22	70.98
SYNTH-343	83.83 $\pm$ 1.61	79.00 $\pm$ 1.80	61.67 $\pm$ 2.08	61.50 $\pm$ 3.61	85.02	61.02	73.02
SYNTH-344	86.33 $\pm$ 4.31	80.67 $\pm$ 2.47	65.00 $\pm$ 3.61	62.00 $\pm$ 1.00	86.10	59.97	73.03
SYNTH-345	88.50 $\pm$ 1.80	81.83 $\pm$ 5.35	67.67 $\pm$ 5.86	75.50 $\pm$ 6.50	87.22	64.72	75.97
SYNTH-346	89.67 $\pm$ 4.51	82.00 $\pm$ 2.29	72.33 $\pm$ 3.33	68.50 $\pm$ 3.04	86.38	66.67	76.52
SYNTH-347	91.33 $\pm$ 2.36	84.83 $\pm$ 1.53	74.83 $\pm$ 4.86	75.33 $\pm$ 4.86	87.65	69.83	78.74
SYNTH-348	88.67 $\pm$ 3.25	81.00 $\pm$ 3.77	73.33 $\pm$ 3.21	72.33 $\pm$ 7.22	85.38	71.52	78.45
SYNTH-349	88.50 $\pm$ 1.73	81.50 $\pm$ 2.00	75.50 $\pm$ 1.00	80.00 $\pm$ 1.73	83.05	75.45	79.25
SYNTH-350	89.17 $\pm$ 1.76	81.33 $\pm$ 1.04	77.33 $\pm$ 2.57	77.67 $\pm$ 3.55	87.77	77.08	82.43
SYNTH-351	92.33 $\pm$ 1.26	87.00 $\pm$ 4.77	81.33 $\pm$ 3.33	82.00 $\pm$ 2.29	84.73	81.47	83.10
SYNTH-352	91.17 $\pm$ 3.21	84.50 $\pm$ 3.12	83.67 $\pm$ 3.69	80.00 $\pm$ 0.50	85.33	81.20	83.27
SYNTH-353	91.17 $\pm$ 2.57	89.00 $\pm$ 3.12	88.17 $\pm$ 3.40	85.83 $\pm$ 2.02	87.02	85.75	86.38
SYNTH-354	93.67 $\pm$ 2.25	88.00 $\pm$ 3.00	85.67 $\pm$ 3.55	85.67 $\pm$ 2.75	85.77	85.17	85.47
SYNTH-355	90.33 $\pm$ 2.08	90.33 $\pm$ 1.76	89.17 $\pm$ 3.01	90.33 $\pm$ 1.89	83.23	90.82	87.03
SYNTH-356	93.00 $\pm$ 1.80	89.67 $\pm$ 2.36	90.83 $\pm$ 1.04	88.67 $\pm$ 1.61	85.42	89.30	87.36
SYNTH-357	95.17 $\pm$ 1.26	94.67 $\pm$ 0.58	94.00 $\pm$ 1.00	94.00 $\pm$ 1.80	84.82	96.17	90.49
SYNTH-358	96.00 $\pm$ 1.00	95.17 $\pm$ 1.26	95.67 $\pm$ 1.53	95.33 $\pm$ 2.36	83.42	94.03	88.72
SYNTH-359	98.17 $\pm$ 0.58	98.50 $\pm$ 0.87	98.33 $\pm$ 0.29	97.83 $\pm$ 0.76	87.90	98.38	93.14
SYNTH-360	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	83.57	100.00	91.78
SYNTH-361	86.00 $\pm$ 2.50	78.00 $\pm$ 1.32	55.50 $\pm$ 4.77	40.17 $\pm$ 4.31	88.45	32.47	60.46
SYNTH-362	83.83 $\pm$ 1.89	78.83 $\pm$ 3.01	59.50 $\pm$ 4.36	42.33 $\pm$ 2.57	90.22	34.25	62.23
SYNTH-363	91.00 $\pm$ 2.29	85.50 $\pm$ 3.28	64.67 $\pm$ 5.58	38.67 $\pm$ 3.01	92.58	37.50	65.04
SYNTH-364	89.33 $\pm$ 2.02	84.00 $\pm$ 0.87	64.50 $\pm$ 5.00	47.00 $\pm$ 3.91	92.77	36.18	64.47
SYNTH-365	89.17 $\pm$ 1.76	82.33 $\pm$ 2.89	61.17 $\pm$ 6.64	48.33 $\pm$ 5.53	90.60	43.83	67.22
SYNTH-366	84.00 $\pm$ 0.50	79.00 $\pm$ 4.58	59.00 $\pm$ 7.79	48.17 $\pm$ 2.25	88.10	45.72	66.91
SYNTH-367	86.33 $\pm$ 2.75	81.00 $\pm$ 1.00	64.83 $\pm$ 8.78	51.00 $\pm$ 3.28	89.82	45.58	67.70
SYNTH-368	85.17 $\pm$ 2.31	83.33 $\pm$ 1.61	63.67 $\pm$ 4.62	48.33 $\pm$ 5.35	89.28	45.53	67.41
SYNTH-369	89.00 $\pm$ 2.50	86.83 $\pm$ 3.33	70.83 $\pm$ 5.06	57.50 $\pm$ 0.87	91.80	55.73	73.77
SYNTH-370	89.83 $\pm$ 1.89	87.00 $\pm$ 2.65	69.67 $\pm$ 8.69	55.17 $\pm$ 5.48	90.07	52.73	71.40
SYNTH-371	89.00 $\pm$ 0.50	84.67 $\pm$ 0.76	70.17 $\pm$ 5.25	67.50 $\pm$ 3.50	92.33	54.87	73.60
SYNTH-372	88.33 $\pm$ 1.89	83.33 $\pm$ 1.15	69.83 $\pm$ 3.55	62.83 $\pm$ 4.07	90.27	59.52	74.89
SYNTH-373	91.50 $\pm$ 0.50	88.00 $\pm$ 1.50	77.00 $\pm$ 5.63	67.17 $\pm$ 2.47	93.65	59.80	76.72
SYNTH-374	89.33 $\pm$ 1.44	81.33 $\pm$ 2.47	70.00 $\pm$ 1.80	64.83 $\pm$ 4.54	86.97	63.42	75.19
SYNTH-375	90.67 $\pm$ 1.53	83.83 $\pm$ 1.76	71.17 $\pm$ 4.16	67.17 $\pm$ 3.79	88.42	66.03	77.22
SYNTH-376	92.50 $\pm$ 0.50	88.00 $\pm$ 3.50	76.00 $\pm$ 2.50	76.83 $\pm$ 0.76	89.87	66.87	78.37
SYNTH-377	93.00 $\pm$ 2.65	85.33 $\pm$ 3.33	76.83 $\pm$ 5.11	74.33 $\pm$ 5.62	90.03	73.03	81.53
SYNTH-378	92.67 $\pm$ 2.31	87.33 $\pm$ 1.89	75.50 $\pm$ 3.28	71.50 $\pm$ 5.27	92.07	68.53	80.30
SYNTH-379	92.67 $\pm$ 1.04	88.83 $\pm$ 1.61	80.50 $\pm$ 2.65	79.67 $\pm$ 2.08	92.93	74.77	83.85
SYNTH-380	93.83 $\pm$ 1.26	89.17 $\pm$ 0.29	75.83 $\pm$ 2.57	83.00 $\pm$ 2.65	91.47	75.12	83.29
SYNTH-381	90.33 $\pm$ 1.26	81.50 $\pm$ 3.04	78.83 $\pm$ 0.58	76.33 $\pm$ 0.76	90.03	79.42	84.73
SYNTH-382	93.83 $\pm$ 1.53	89.50 $\pm$ 1.32	83.33 $\pm$ 2.31	85.00 $\pm$ 2.18	91.78	80.42	86.10
SYNTH-383	93.33 $\pm$ 2.08	88.00 $\pm$ 2.18	86.67 $\pm$ 1.61	85.33 $\pm$ 1.26	90.27	84.42	87.34
SYNTH-384	93.33 $\pm$ 2.84	90.50 $\pm$ 2.29	89.50 $\pm$ 0.50	88.50 $\pm$ 2.18	89.83	87.08	88.46
SYNTH-385	95.33 $\pm$ 0.76	90.50 $\pm$ 1.80	90.33 $\pm$ 2.57	88.33 $\pm$ 2.52	89.83	87.62	88.72
SYNTH-386	96.83 $\pm$ 0.76	94.33 $\pm$ 1.04	93.50 $\pm$ 1.50	93.00 $\pm$ 1.00	88.17	90.70	89.43
SYNTH-387	98.00 $\pm$ 0.50	97.33 $\pm$ 0.58	96.67 $\pm$ 0.29	97.33 $\pm$ 1.26	93.15	96.87	95.01
SYNTH-388	95.67 $\pm$ 0.76	93.67 $\pm$ 0.76	93.17 $\pm$ 1.26	95.00 $\pm$ 0.87	90.85	93.58	92.22
SYNTH-389	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	90.65	100.00	95.32
SYNTH-390	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	90.73	100.00	95.37
SYNTH-391	92.33 $\pm$ 2.57	88.67 $\pm$ 2.57	67.67 $\pm$ 3.25	44.67 $\pm$ 7.52	96.37	34.38	65.37
SYNTH-392	90.83 $\pm$ 2.25	88.50 $\pm$ 1.80	69.83 $\pm$ 3.55	47.50 $\pm$ 3.12	94.63	32.07	63.35
SYNTH-393	92.67 $\pm$ 1.76	88.67 $\pm$ 2.75	66.83 $\pm$ 6.60	44.33 $\pm$ 3.79	95.85	39.45	67.65
SYNTH-394	92.33 $\pm$ 3.01	87.67 $\pm$ 0.29	66.33 $\pm$ 5.97	48.67 $\pm$ 1.76	95.70	36.15	65.92

CONTINUED ON NEXT PAGE

Table 18: Synthetic sweep results reported in dataset-style format. All GNN results are reported as mean $\pm$ standard deviation (in percentage points), computed over multiple random seeds. Synthetic datasets were generated consistently across backbones using 1000 graphs per configuration, approximately 120 nodes (± 20 variance), feature dimension 16, and a 15×15 SLA/FLA grid. AlignAvg denotes the arithmetic mean of SLA and FLA (lower is worse).

DATASET	GCN	SAGE	GAT	GPS	SLA	FLA	ALIGNAVG
SYNTH-395	90.83 $\pm$ 2.31	88.83 $\pm$ 0.29	73.33 $\pm$ 4.65	49.33 $\pm$ 2.08	94.92	44.32	69.62
SYNTH-396	91.00 $\pm$ 1.50	87.17 $\pm$ 2.75	70.17 $\pm$ 3.40	50.50 $\pm$ 2.00	96.03	40.62	68.32
SYNTH-397	92.83 $\pm$ 1.04	87.50 $\pm$ 0.50	69.67 $\pm$ 2.25	53.50 $\pm$ 1.32	94.90	49.38	72.14
SYNTH-398	91.33 $\pm$ 0.29	87.83 $\pm$ 0.58	73.17 $\pm$ 3.75	55.33 $\pm$ 8.52	94.73	47.32	71.02
SYNTH-399	89.17 $\pm$ 2.25	87.50 $\pm$ 2.78	73.67 $\pm$ 2.75	58.67 $\pm$ 4.01	94.53	50.73	72.63
SYNTH-400	90.17 $\pm$ 3.33	87.67 $\pm$ 3.75	75.17 $\pm$ 5.13	58.50 $\pm$ 3.04	95.57	54.45	75.01
SYNTH-401	93.33 $\pm$ 1.15	88.83 $\pm$ 1.53	68.00 $\pm$ 6.87	66.33 $\pm$ 7.52	93.93	55.25	74.59
SYNTH-402	95.33 $\pm$ 2.08	91.83 $\pm$ 2.93	78.83 $\pm$ 2.52	67.17 $\pm$ 3.69	96.08	59.02	77.55
SYNTH-403	93.33 $\pm$ 2.84	91.67 $\pm$ 2.52	78.67 $\pm$ 0.76	69.50 $\pm$ 5.57	96.92	62.98	79.95
SYNTH-404	90.83 $\pm$ 1.04	89.00 $\pm$ 1.32	73.17 $\pm$ 7.65	62.67 $\pm$ 2.47	93.58	60.63	77.11
SYNTH-405	91.50 $\pm$ 0.87	88.83 $\pm$ 3.33	79.67 $\pm$ 4.07	79.33 $\pm$ 4.19	95.62	70.57	83.09
SYNTH-406	91.33 $\pm$ 0.76	88.17 $\pm$ 1.61	77.33 $\pm$ 4.65	76.17 $\pm$ 1.44	94.42	70.58	82.50
SYNTH-407	92.33 $\pm$ 2.25	88.50 $\pm$ 2.00	77.17 $\pm$ 0.29	73.00 $\pm$ 1.80	95.75	69.93	82.84
SYNTH-408	93.50 $\pm$ 1.80	91.50 $\pm$ 1.00	82.83 $\pm$ 1.61	82.67 $\pm$ 0.29	96.33	70.18	83.26
SYNTH-409	95.67 $\pm$ 1.53	89.67 $\pm$ 1.53	84.67 $\pm$ 1.44	82.50 $\pm$ 6.24	96.27	78.47	87.37
SYNTH-410	93.17 $\pm$ 1.89	92.83 $\pm$ 0.58	87.33 $\pm$ 0.29	81.50 $\pm$ 8.32	94.85	78.32	86.58
SYNTH-411	93.00 $\pm$ 3.91	90.00 $\pm$ 5.20	80.50 $\pm$ 5.07	83.00 $\pm$ 2.29	93.73	81.93	87.83
SYNTH-412	94.17 $\pm$ 2.08	89.33 $\pm$ 2.52	85.83 $\pm$ 2.02	84.50 $\pm$ 1.00	96.07	80.52	88.29
SYNTH-413	96.17 $\pm$ 0.76	93.00 $\pm$ 3.12	89.00 $\pm$ 1.32	86.67 $\pm$ 0.76	98.12	85.32	91.72
SYNTH-414	94.33 $\pm$ 1.61	90.50 $\pm$ 0.87	84.00 $\pm$ 0.50	83.67 $\pm$ 0.76	94.42	83.68	89.05
SYNTH-415	96.17 $\pm$ 0.76	94.33 $\pm$ 0.29	91.17 $\pm$ 1.44	90.67 $\pm$ 2.02	95.25	90.53	92.89
SYNTH-416	95.00 $\pm$ 3.00	95.50 $\pm$ 1.32	94.17 $\pm$ 1.61	91.83 $\pm$ 2.75	95.23	89.68	92.46
SYNTH-417	98.33 $\pm$ 1.04	98.00 $\pm$ 1.00	97.50 $\pm$ 1.50	98.17 $\pm$ 1.04	94.90	96.80	95.85
SYNTH-418	96.67 $\pm$ 1.15	94.33 $\pm$ 0.76	94.33 $\pm$ 1.15	94.33 $\pm$ 1.04	95.53	94.20	94.87
SYNTH-419	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	96.72	100.00	98.36
SYNTH-420	99.17 $\pm$ 0.58	99.00 $\pm$ 0.50	99.00 $\pm$ 0.50	99.33 $\pm$ 0.29	95.70	99.45	97.58
SYNTH-421	94.83 $\pm$ 2.47	95.83 $\pm$ 1.53	85.50 $\pm$ 2.18	49.17 $\pm$ 6.03	100.00	33.63	66.82
SYNTH-422	96.67 $\pm$ 1.26	95.50 $\pm$ 0.87	86.17 $\pm$ 4.65	48.17 $\pm$ 4.86	100.00	34.52	67.26
SYNTH-423	94.33 $\pm$ 2.93	92.83 $\pm$ 3.33	83.83 $\pm$ 3.06	50.67 $\pm$ 5.84	100.00	37.52	68.76
SYNTH-424	95.33 $\pm$ 1.53	94.33 $\pm$ 2.08	84.50 $\pm$ 3.61	49.67 $\pm$ 4.51	99.67	39.58	69.62
SYNTH-425	93.50 $\pm$ 2.65	92.17 $\pm$ 3.01	83.17 $\pm$ 5.35	48.83 $\pm$ 7.29	100.00	42.13	71.07
SYNTH-426	94.67 $\pm$ 2.31	92.83 $\pm$ 2.25	84.67 $\pm$ 1.04	44.67 $\pm$ 5.13	100.00	41.37	70.68
SYNTH-427	93.83 $\pm$ 2.57	90.50 $\pm$ 2.00	75.17 $\pm$ 4.75	55.50 $\pm$ 6.87	97.68	45.30	71.49
SYNTH-428	93.67 $\pm$ 1.26	93.83 $\pm$ 2.02	82.50 $\pm$ 3.50	54.50 $\pm$ 3.97	100.00	45.75	72.88
SYNTH-429	95.17 $\pm$ 1.53	94.50 $\pm$ 1.00	87.00 $\pm$ 2.00	65.50 $\pm$ 1.32	99.07	53.47	76.27
SYNTH-430	94.33 $\pm$ 1.15	93.00 $\pm$ 2.60	82.50 $\pm$ 5.89	59.50 $\pm$ 7.00	98.07	51.65	74.86
SYNTH-431	95.67 $\pm$ 1.53	96.17 $\pm$ 1.53	87.50 $\pm$ 3.04	70.00 $\pm$ 12.22	100.00	55.90	77.95
SYNTH-432	94.17 $\pm$ 1.61	93.33 $\pm$ 1.61	83.17 $\pm$ 2.02	65.67 $\pm$ 3.40	99.45	54.57	77.01
SYNTH-433	96.17 $\pm$ 0.76	95.33 $\pm$ 1.53	87.33 $\pm$ 1.53	74.17 $\pm$ 4.16	100.00	62.55	81.28
SYNTH-434	96.83 $\pm$ 1.15	93.50 $\pm$ 1.50	84.17 $\pm$ 1.26	73.33 $\pm$ 4.04	99.07	67.93	83.50
SYNTH-435	95.83 $\pm$ 2.02	95.50 $\pm$ 1.80	87.33 $\pm$ 4.48	75.33 $\pm$ 6.11	100.00	65.83	82.92
SYNTH-436	94.33 $\pm$ 1.61	93.83 $\pm$ 1.53	86.67 $\pm$ 3.33	80.17 $\pm$ 7.32	99.50	66.75	83.12
SYNTH-437	96.50 $\pm$ 1.80	94.33 $\pm$ 2.93	86.17 $\pm$ 2.25	79.33 $\pm$ 2.02	100.00	71.08	85.54
SYNTH-438	93.33 $\pm$ 2.47	93.33 $\pm$ 2.57	86.67 $\pm$ 3.18	84.00 $\pm$ 1.73	100.00	71.58	85.79
SYNTH-439	95.00 $\pm$ 1.80	94.17 $\pm$ 2.02	89.00 $\pm$ 0.50	86.83 $\pm$ 1.76	98.73	77.43	88.08
SYNTH-440	95.67 $\pm$ 1.61	92.83 $\pm$ 1.61	85.83 $\pm$ 2.02	76.00 $\pm$ 1.00	99.35	72.83	86.09
SYNTH-441	96.33 $\pm$ 1.76	94.33 $\pm$ 2.84	90.33 $\pm$ 2.25	85.83 $\pm$ 1.76	100.00	80.72	90.36
SYNTH-442	95.83 $\pm$ 2.08	93.67 $\pm$ 0.29	90.17 $\pm$ 1.44	85.50 $\pm$ 4.44	100.00	78.53	89.27
SYNTH-443	96.83 $\pm$ 1.15	93.83 $\pm$ 2.08	93.17 $\pm$ 0.58	91.00 $\pm$ 0.50	99.77	88.68	94.22
SYNTH-444	97.33 $\pm$ 0.29	94.83 $\pm$ 2.75	90.33 $\pm$ 1.26	88.67 $\pm$ 0.29	100.00	86.07	93.03
SYNTH-445	97.50 $\pm$ 0.50	91.00 $\pm$ 1.32	91.83 $\pm$ 1.89	89.33 $\pm$ 1.61	100.00	91.98	95.99
SYNTH-446	95.67 $\pm$ 0.29	91.17 $\pm$ 2.75	93.17 $\pm$ 0.29	91.17 $\pm$ 0.76	99.20	90.97	95.08
SYNTH-447	95.83 $\pm$ 1.26	94.33 $\pm$ 0.76	94.00 $\pm$ 0.50	94.67 $\pm$ 0.29	100.00	95.10	97.55
SYNTH-448	97.50 $\pm$ 0.87	96.50 $\pm$ 0.87	96.50 $\pm$ 0.50	96.33 $\pm$ 1.15	100.00	95.90	97.95
SYNTH-449	99.17 $\pm$ 0.29	99.67 $\pm$ 0.58	99.33 $\pm$ 0.58	99.33 $\pm$ 0.29	99.05	99.12	99.08
SYNTH-450	99.33 $\pm$ 0.29	99.33 $\pm$ 0.29	99.33 $\pm$ 0.58	99.50 $\pm$ 0.00	98.57	99.43	99.00

Table 19: Wall-clock seconds for one pass of GLA (one seed, 10-fold CV, LogReg and RBF-SVM on SLA and FLA) against one 100-epoch training run of each backbone with a single hyperparameter configuration, on one machine.

DATASET	GRAPHS	SLA BUILD	SLA CV	FLA BUILD	FLA CV	GLA	GCN	SAGE	GAT	GPS	SUM	RATIO
MUTAG	188	0.30	0.05	0.01	0.03	0.39	1.32	1.02	2.92	4.15	9.4	23.9×
BZR	405	1.32	0.12	0.02	0.10	1.55	3.18	2.69	10.72	14.49	31.1	20.1×
PROTEINS	1,113	4.40	0.55	0.02	0.13	5.11	10.13	7.72	36.46	82.27	136.6	26.7×
NCI1	4,110	12.13	3.27	0.09	3.73	19.23	32.18	28.62	137.29	152.75	350.8	18.2×

Table 20: Leave-one-dataset-out prediction of GCN accuracy from SLA, FLA, and AlignAvg with a linear model, on the 19 benchmarks where FLA is defined, sorted by absolute residual (observed minus predicted).

DATASET	OBS.	PRED.	GAP	DATASET	OBS.	PRED.	GAP
SYNTHIE	69.0	53.0	+16.0	DD	70.3	75.2	-4.9
COX2	78.1	65.0	+13.1	PROTEINS	67.6	71.0	-3.4
MOLTOX21	73.5	86.0	-12.6	PROTEINS_FULL	71.1	74.3	-3.3
COLORS-3	24.9	35.8	-10.9	BACE	72.8	70.4	+2.3
CIFAR-10	27.8	36.8	-9.0	LETTER-HIGH	41.4	43.4	-2.0
MUTAG	71.3	79.7	-8.4	ENZYMES	47.5	46.0	+1.6
AIDS	98.4	90.4	+8.0	NCI1	68.3	67.2	+1.1
COIL-DEL	34.0	26.2	+7.8	MOLHIV	71.4	72.0	-0.6
BZR	81.5	74.2	+7.3	MNIST	37.9	38.0	-0.1
FRANKENSTEIN	59.1	66.3	-7.1				

H Screening Use and Split Protocol

This appendix evaluates alignment as a pre-training screen: what it costs, how well it predicts accuracy on unseen datasets, and what it says about benchmark quality. It then tests how far the results extend beyond random splits.

H.1 Compute Cost

Table 19 compares one pass of the alignment pipeline with one training run of each backbone, across a 22-fold range of dataset sizes on a single CPU-only machine.

GLA costs 18–27 times less than one training run of all four backbones, and 8–16 times less than GPS alone. Structural summary construction takes most of the GLA budget (63–86%), while the centroid is nearly free.

These ratios should be read as a measured comparison of these specific runs, not as a general speedup. Two choices pull in opposite directions. On the GLA side, we time one seed, whereas the reference protocol uses five. Multiplying only the probe columns by five gives an estimated ratio of 7–17. On the GNN side, we time a single configuration on a single split, with no evaluation, whereas our grid search (Section A) has 12 configurations for GCN and SAGE and 48 for GAT and GPS. A GPU would shorten training and narrow the ratio. Our largest test has 4,110 graphs, and we do not extrapolate beyond it: the probes’ share of GLA is largest on NCI1 (36%), and the RBF-SVM scales superlinearly in the number of graphs (Section B.4), so linear probes may be needed at larger scale.

H.2 Out-of-Sample Prediction

The per-backbone predictors in Section G are not tested on held-out datasets. We therefore run leave-one-dataset-out cross-validation: a model fitted on 18 benchmarks predicts GCN accuracy on the 19th from its alignment scores alone. A linear model reaches Pearson $r = 0.917$, Spearman $\rho = 0.821$, MAE 6.30, and RMSE 7.77. Ridge regression gives the same result, and a random forest does worse ($r = 0.884$, RMSE 9.18). Since AlignAvg is a linear combination of SLA and FLA, the linear model effectively uses two inputs.

Table 20 reports the residual for every dataset. The eight residuals larger than the RMSE are the cases worth inspecting. Positive gaps (Synthie, COX2, AIDS, COIL-DEL) mark datasets where GCN exceeds what alignment predicts, and negative gaps (MOLTOX21, COLORS-3, CIFAR-10, MUTAG) mark datasets where it falls short. With 19 datasets and a typical error of 6–8 points, this supports ranking datasets and flagging large departures. It does not support precise accuracy estimates, and we

Table 21: Δ_{graph} (accuracy points) for the 19 benchmarks where FLA is defined, computed from Table 2.

$\Delta_{\text{graph}} \leq +2$		$+2 < \Delta_{\text{graph}} < +10$		$\Delta_{\text{graph}} \geq +10$	
MUTAG	-15.4	BACE	+5.8	COX2	+11.4
MOLTOX21	-15.0	BZR	+6.2	ENZYMES	+14.1
FRANKENSTEIN	-5.9			CIFAR-10	+17.6
DD	-4.9			MNIST	+23.0
PROTEINS	-4.6			SYNTHIE	+29.7
MOLHIV	-2.2			COIL-DEL	+37.8
AIDS	-0.7			LETTER-HIGH	+43.3
PROTEINS_FULLL	+0.1				
NCI1	+0.9				
COLORS-3	+1.9				

have tested it only for GCN. This suggests a three-step workflow: compute SLA and FLA, predict accuracy with a model fitted on other datasets, and spend training budget where the prediction or the SLA-FLA gap warrants it.

H.3 Benchmark Triage

Following the call to validate graph benchmarks against baselines that ignore relational structure [BSFF⁺25, CWSR25], we define

$$\Delta_{\text{graph}} = \text{Acc}(\text{best pooled backbone}) - \max(\text{SLA}, \text{FLA}),$$

the accuracy a trained graph model adds over the best no-graph probe. It requires no training beyond the baselines in Table 2. The restricted views correspond to the modes in RINGS: the feature-only view discards the graph and the structure-only view discards features. RINGS estimates these modes by retraining GNNs, whereas we use shallow probes, so Δ_{graph} can act as a cheap screen before perturbation-based evaluation. Since our probes are deliberately simple, Δ_{graph} is an upper bound on what relational modeling adds under this protocol, and it depends on how well the backbones were tuned. Negative values mean a shallow probe beat every backbone we trained.

Ten of the 19 benchmarks have $\Delta_{\text{graph}} \leq +2$, and a shallow probe beats the best backbone on seven of them. Nine of the ten are molecular or bioinformatics benchmarks. When accuracy is also high, as on AIDS, these datasets cannot separate methods that use relational structure from methods that discard it, and they should not anchor headline GNN comparisons on their own. Datasets with $\Delta_{\text{graph}} \geq +10$, largely vision-derived or synthetic, are those where trained graph models add accuracy that neither marginal view reaches.

Δ_{graph} must be read together with absolute accuracy, because the gap can also vanish when nothing solves the task. COLORS-3 ($\Delta_{\text{graph}} = +1.9$ at 26% best accuracy) and the Twins are of this kind. They are hard rather than saturated, and they belong in benchmark suites. Δ_{graph} is undefined on the featureless social benchmarks.

H.4 Scaffold Splits

Our real-benchmark protocol uses random splits, so it measures in-distribution accessibility. Molecular evaluation often uses scaffold splits, which hold out molecular scaffolds and so test generalization under distribution shift. We recompute SLA and FLA and train a GCN under the official OGB scaffold split, and compare with a random split of the same train and test sizes, which isolates the split type from the split size (Table 22). We report balanced accuracy and ROC-AUC because plain accuracy is uninformative on the imbalanced datasets.

Scaffold splits change the picture on most of these datasets. By ROC-AUC, GCN falls on five of the six, all except MOLTOX21. SLA falls on all six, though only slightly on MOLHIV, and FLA falls on four, staying flat on BACE and MOLTOX21. The three scores also agree on where the effect is largest: BBBP shows the largest drop for all of them (-0.20 for SLA and FLA, -0.28 for GCN), followed by ClinTox. Alignment therefore tends to weaken along with GCN under this shift, but it does not track GCN dataset by dataset, as MOLTOX21 shows.

We read these results as a caution about scope. Our random-split correlations do not establish performance under distribution shift, and a practitioner using scaffold splits should compute alignment

Table 22: Official OGB scaffold split (S) against a random split of the same train and test sizes (R). SLA and FLA average LogReg and RBF-SVM, and R averages 5 seeds. GCN is a single untuned run on each split with unweighted cross-entropy. BA is balanced accuracy and AUC is ROC-AUC, using the first task on multi-task datasets.

DATASET	SLA				FLA				GCN			
	BA		AUC		BA		AUC		BA		AUC	
	S	R	S	R	S	R	S	R	S	R	S	R
BACE	0.66	0.72	0.75	0.79	0.62	0.65	0.73	0.73	0.54	0.70	0.78	0.81
MOLTOX21	0.64	0.73	0.74	0.79	0.70	0.73	0.80	0.80	0.68	0.58	0.76	0.72
MOLHIV	0.65	0.65	0.70	0.71	0.63	0.63	0.64	0.67	0.50	0.50	0.62	0.69
BBBP	0.56	0.72	0.60	0.80	0.58	0.76	0.64	0.83	0.52	0.76	0.65	0.93
CLINTOX	0.49	0.52	0.53	0.62	0.54	0.66	0.58	0.71	0.50	0.50	0.61	0.75
SIDER	0.58	0.58	0.58	0.62	0.57	0.59	0.56	0.64	0.55	0.61	0.57	0.61

on those splits. The evidence is also limited. GCN is a single untuned run per split, and its unweighted loss differs from the protocol behind Table 2. On MOLHIV and ClinTox it collapses to the majority class on both splits (balanced accuracy near 0.50), so only ROC-AUC is informative there. Test sets on the smaller datasets have only 143–204 molecules, so small differences should not be over-read.

I Limitations

Our framework is intended as a diagnostic rather than a replacement for full model evaluation. SLA and FLA measure label accessibility under fixed structure-only and feature-only summaries, so their values are relative to the chosen summaries and probes.

Summary and probe dependence. Low alignment on real data may reflect weak marginal signal, interaction-dependent signal, or limits of the summaries and probes. Richer feature summaries and stronger probes raise scores on some low-alignment datasets (Sections C.2 and C.3), so such scores are best read as lower bounds. Twins are the exception: because their marginals are matched across classes by construction, no summary of either restricted view can separate them. Rankings and joint regimes are stable across the fixed summaries and probes we tested (Section C.4), but we did not evaluate trained summaries such as Deep Sets, and we varied structural blocks without varying the number of eigenvalues or the filtration resolution within a block.

Observational and interventional evidence. Correlations on real benchmarks are descriptive, since domain, size, features, and class count vary together. Evidence that alignment affects learnability comes from the synthetic sweep, where structural and feature separation are set directly while other factors are held fixed, and from Twins.

Split protocol. Our real-benchmark correlations are an in-distribution diagnostic. Seed and split stability holds for random splits (Section C.5), while scaffold splits can change the picture (Section H.4). We evaluated scaffold splits on molecular benchmarks only, and other domain-aware protocols remain open.

Scope. The framework requires a label that belongs to the whole graph, so that discarding edges or features is a well-defined intervention. Our ego-graph experiments recast node classification in this form and show where the diagnostic weakens (Section D). Graph regression should need only regressors in place of the probes, though we have not tested this. Link prediction is harder: its label belongs to a node pair and structure is partly what is being predicted, so its restricted views would need to be redefined. Our study also focuses on standard pooled classifiers and benchmarks of limited size, with screening cost measured up to 4,110 graphs (Section H.1).

These limitations do not affect the main conclusion: graph label alignment provides a lightweight, architecture-agnostic way to organize benchmark difficulty and identify when standard pooled pipelines are likely to succeed or struggle.

J Broader Impacts

This paper introduces graph label alignment, a probe based diagnostic that estimates where label signal resides in graph classification datasets by measuring structural and feature marginal accessibility via SLA and FLA. The main positive impact is improved transparency and rigor in benchmark driven graph learning: researchers can better interpret reported gains, identify when high accuracy is likely due to coarse shortcuts, and design evaluations that meaningfully test relational reasoning, which can ultimately improve model reliability in scientific applications such as chemistry and biology. The proposed alignment plane and controlled testbeds may also help practitioners choose appropriate model families and avoid unnecessary compute by revealing when pooled architectures are unlikely to succeed. Potential negative impacts include misuse of these diagnostics to engineer datasets or leaderboards that favor specific methods, or to over simplify conclusions by treating SLA and FLA as definitive measures of task difficulty rather than informative probes. In addition, applying the framework to sensitive relational datasets could facilitate more effective classification of individuals or groups, with attendant privacy and fairness concerns. We encourage responsible use by reporting alignment scores alongside standard metrics, validating conclusions across multiple descriptors and probes, auditing for subgroup harms when labels permit, and avoiding deployment of graph classifiers on high stakes social data without careful governance and monitoring.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes].

Justification: The claims match the proposed SLA/FLA diagnostic framework, the real and synthetic experiments, and the stated scope as a diagnostic rather than a causal model.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes].

Justification: The paper discusses limits of restricted-view probes, probe sensitivity, computational costs, and the scope of the evaluated pooled graph classifiers (Appendix I).

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper is primarily empirical and diagnostic, and does not present formal theoretical results requiring proofs.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?

Answer: [Yes].

Justification: The paper specifies datasets, splits, models, probes, hyperparameter grids, synthetic generation details, and evaluation protocols in the main text and appendix.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: The paper provides an anonymized code link and uses public benchmark datasets together with documented synthetic data generation.

6. Experimental setting/details

Question: Does the paper specify all the training and test details necessary to understand the results?

Answer: [Yes].

Justification: Training splits, model families, pooling choices, probe protocols, seeds, cross-validation, and hyperparameter grids are described in the experimental section and appendix.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes].

Justification: Main benchmark tables report mean accuracy with standard deviations, and the alignment correlations include statistical significance levels.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

Answer: [Yes].

Justification: The probes are lightweight, and the paper describes algorithmic complexity and scalability.

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

Answer: [Yes].

Justification: The work uses public datasets and synthetic data, does not involve human subjects, and is intended for diagnostic analysis of graph learning benchmarks.

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: The broader impacts discussion addresses improved benchmark transparency and notes that the work is diagnostic, with no direct deployment pathway (Appendix J).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

Answer: [N/A].

Justification: The paper does not release high-risk models, scraped datasets, or dual-use generative systems.

12. Licenses for existing assets

Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites the benchmark datasets and model components used; licenses and terms should also be included in the released code documentation.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: The paper documents the synthetic datasets, alignment probes, and code needed to reproduce the proposed diagnostic framework.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject experiments.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether IRB approvals were obtained?

Answer: [N/A].

Justification: The paper does not involve human subjects, so IRB approval is not applicable.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A].

Justification: LLMs are not part of the core method, experiments, or scientific contribution.