MSS: MULTI-SEMANTIC SPACES AS A DECLARED EVIDENCE LAYER FOR GRAPH ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmarks for graph anomaly detection (GAD) usually change the evidence construction and the consuming model together, so a reported margin does not show which of the two produced it, and an aggregate score hides the view a prediction rests on. Holding the read-out fixed locates the gain in the construction rather than in the model that consumes it: on the same tree head, complete evidence improves the ten-graph mean AUPRC over its baseline feed by 1.65 points under full supervision and 4.17 under the 20+80 budget, and on the largest graph adding regional references to that feed moves AUPRC from 41.12 to 72.64. We propose Multi-Semantic Spaces (MSS), a framework that treats raw and graph-derived feature blocks as *materials* declared by explicit mechanism hypotheses. A frozen topology backbone, soft regional occupancy, and regional references organize these materials into positions and deviations from a declared comparison unit. MSS-T fits boosted trees to a fixed evidence table, while MSS-N learns private coordinates and per-space experts, updates its references during training, and retains per-space states through stop-gradient fusion. Complete evidence also improves a second library-default head. On ten public benchmarks, using the released GADBench splits and a common ten-seed list, one configuration shared across all ten datasets and never retuned attains the highest mean AUPRC under both budgets. It exceeds the strongest baseline by 2.57 points under full supervision and 4.69 under the 20+80 budget. MSS-N exceeds MSS-T on two fully supervised graphs and exposes coordinate-level diagnostics, so the gains stay attributable and the states inspectable.

1 INTRODUCTION

Graph anomaly detection (GAD) seeks nodes, edges, or subgraphs that depart from the prevailing patterns of a graph (Liu et al., 2022; Tang et al., 2023). Applications include financial fraud and money laundering (Huang et al., 2022; Weber et al., 2019), review spam (Rayana & Akoglu, 2015; McAuley & Leskovec, 2013), and coordinated accounts (Beutel et al., 2013; Cao et al., 2014). Anomalies rarely share one surface pattern. A collusive group shows repeated interactions, a disposable account shows sparse and short-lived connectivity, a fabricated profile shows attribute deviation, and a money mule shows structural role, temporal rhythm, or edge direction. These mechanisms can coexist under severe class imbalance, heavy-tailed features, and relational camouflage (Tang et al., 2023; Dou et al., 2020). Consequently, a representation assumption that works for ordinary node classification is not necessarily suitable for GAD. Existing detectors follow several complementary routes: spectral and heterophily-aware designs (Tang et al., 2022; Gao et al., 2023; Shi et al., 2022; Xu et al., 2024), tree ensembles over graph statistics (Breiman, 2001; Chen & Guestrin, 2016; Tang et al., 2023), and multi-space or label-efficient detectors (Dong et al., 2025; Chen et al., 2024; Qiu et al., 2025). What a node is compared against has rarely been the object of design: prior routes either keep the observed edge neighborhood and change how it is summarized, or learn a reference from geometry or a normality objective (Section 2). Declaring that reference from an anomaly mechanism remains open.

That gap is not hypothetical. GADBench reports that, under common splits, supervision budgets, and execution protocols, many standard GNNs are not stable choices for GAD, while tree ensembles

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

over simple graph statistics are strong by default (Tang et al., 2023). That conclusion rests on a controlled comparison, not on one large margin. Recent methods instead adapt representations to scarce labels and heterogeneous patterns (Ai et al., 2025; Liu et al., 2026b;c). Two gaps remain. First, when evidence construction and the consuming model change together, a benchmark margin alone cannot establish which change helped. We call this an *unattributed performance margin*, and matched controls are needed to separate the effects of materials, their organization, and the read-out. Second, an aggregate score arrives without the view-level state behind it: which material view carried a prediction, and how a node's state within that view changes under the available training labels, are not part of the result. Exposing them requires retaining view-specific positions, scores, and reference states alongside accuracy. Both gaps reduce to one design question: whether the reference a node is compared against can be declared from anomaly mechanisms, whether this evidence improves a fixed read-out, and which view-specific states help inspect its predictions.

Motivated by this question, we propose Multi-Semantic Spaces (MSS), an evidence layer that declares raw and graph-derived feature blocks as *materials* under explicit mechanism hypotheses. Unlike spaces distinguished by embedding geometry (Dong et al., 2025), an MSS space is defined by a mechanism hypothesis and its feature specification. Materials are assembled and normalized over a shared topology backbone, and soft occupancy forms regional references whose deviations carry the node evidence (Figure 1). MSS-T fits boosted trees to the fixed evidence table; MSS-N learns private three-dimensional coordinates and one expert per space, recalibrates its references during training, and fuses expert outputs through stop-gradient gating while retaining per-space positions and scores.

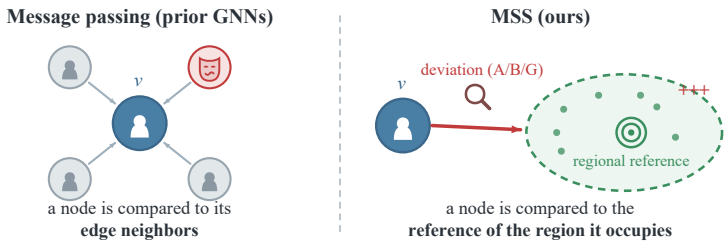

Figure 1: **What a detector reads.** Message passing compares a node against its observed edge neighbors. MSS compares it against a regional reference. Regions are read-out units, not ground-truth communities.

Our contributions are:

- **Declared evidence improves fixed read-outs.** Held at one off-the-shelf tree head, the declared construction improves the ten-graph mean AUPRC by 1.65 and 4.17 points, and adding regional references to the baseline feed moves the largest graph from 41.12 to 72.64 under 20+80 labels. Where message passing compares a node against its observed edge neighbors, MSS compares it against a regional reference declared by a mechanism hypothesis: named materials are assembled over a shared topology backbone, soft occupancy forms regions, and occupancy-conditioned references supply per-node evidence without class-conditioned centers or community labels.
- **One evidence construction, two read-outs.** MSS-T fits boosted trees to the frozen evidence table, and MSS-N learns private coordinates and per-space experts while retaining view-specific scores for diagnosis. The node–region–node read-out avoids dense node-pair matrices, and both modes support declared temporal and directed spaces.
- **Matched controls locate gains and dataset-dependent limits.** Matched removals locate the margin graph by graph, and MSS-T attains the highest mean AUPRC under both budgets (63.77 and 47.86, or +2.57 and +4.69 over the strongest baseline). Dropping the largest graph leaves the ordering intact (59.79 and 44.95). The modes are complementary: MSS-N leads on Reddit and DGraph-Fin under full supervision.

2 Related Work

Benchmarks and message-passing models. BOND benchmarks unsupervised outlier node detection (Liu et al., 2022). GADBench compares standard GNNs, GAD-specific GNNs, classical

models, and tree ensembles under common protocols (Tang et al., 2023). DGraph-Fin exposes the gap to financial graphs with millions of nodes (Huang et al., 2022). Our evaluation inherits the protocol and the million-node scale test. GCN, GraphSAGE, GAT, GIN, Graph Transformer, and PNA propagate through the edge neighborhood (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Tang et al., 2023), and a GAD-specific line (GAS, CARE-GNN, GraphConsis, PC-GNN, GATSep, GTAN, and RGTAN) modifies that local primitive through neighbor selection, camouflage resistance, imbalance handling, and risk-aware representations (Tang et al., 2023; Dou et al., 2020; Liu et al., 2021; Liu et al., 2020; Xiang et al., 2023; 2025). These methods change how information moves along observed edges. MSS changes what a node is compared against.

Spectral, heterophily, and tree-based alternatives. BernNet, AMNet, BWGNN, and GHRN organize anomaly evidence with spectral filters, high-frequency signals, and homophily cues (Tang et al., 2022; Gao et al., 2023; Tang et al., 2023). H2-FDetector and SEC-GFD confirm that heterophily matters in fraud detection (Shi et al., 2022; Xu et al., 2024), and HSMAD adds heterophily-weighted filtering with manifold routing (Zhu & Zhang, 2026). Random Forest, XGBoost, XGBOD, RF-Graph, and XGB-Graph form decision regions over graph-derived features (Breiman, 2001; Chen & Guestrin, 2016; Zhao & Hryniewicki, 2018; Tang et al., 2023) and are strong at default settings. MSS reuses such statistics but supplies the reference that decision regions lack: its materials are declared per mechanism and organized by soft regional occupancy.

Multiple spaces, normality references, and fusion. SpaceGNN learns alternative embedding geometries (Euclidean, hyperbolic, spherical) for limited-label detection (Dong et al., 2025), whereas each MSS space is defined by a mechanism hypothesis. ConsisGAD addresses scarce supervision, while GPA uses path-based aggregation (Chen et al., 2024; Qiu et al., 2025). DeepSVDD, Deep SAD, DPAE, OCGNN, and GGAD build references from normality centers or one-class objectives (Ruff et al., 2018; 2020; Snell et al., 2017; Kang & Buu, 2024; Wang et al., 2021; Qiao et al., 2024), while DiffPool, MinCutPool, and Cluster-GCN learn node assignments for coarsening or minibatching (Ying et al., 2018; Bianchi et al., 2020; Chiang et al., 2019). Such references are derived from geometry or normality objectives, not declared from an anomaly mechanism. Whole-model comparisons alone therefore do not isolate evidence construction from the consuming model. MSS declares the reference per mechanism, computes it once per region, and adds matched tree controls and a neural read-out that keeps per-space states. Appendix A records where each route places its comparison reference.

3 Problem Statement

An attributed graph is $G = (V, E, X)$ with features $X \in \mathbb{R}^{N\times d}$. Node i has feature x_i and receiver-side neighborhood $\mathcal{N}(i)$. Node anomaly detection is binary, $y_i \in \{0, 1\}$ with 1 the anomaly. Only $L \subseteq V$ carries labels under a supervision budget, while a transductive detector may use the graph and features of all nodes. Rare anomalies combine class imbalance with heterogeneous mechanisms. Common GNNs start from $h_i^{(0)} = x_i$ and update a node along observed edges:

$$h_i^{(\ell+1)} = \phi_\ell\Big(h_i^{(\ell)}, \mathrm{AGG}_\ell\Big(\{h_j^{(\ell)} : j \in \mathcal{N}(i)\}, \{e_{ij} : j \in \mathcal{N}(i)\}\Big)\Big). \tag{1}$$

Equation (1) makes the default reference of this line explicit: node i obtains its relational context primarily from $\mathcal{N}(i)$. MSS keeps those materials but takes its reference from a mechanism-specific region.

An anomaly label describes the detection outcome, whereas a mechanism hypothesis describes why a node may be anomalous. The evidence specification $\mathcal{R}_m$ of the m-th mechanism hypothesis $\mathcal{H}_m$ specifies raw or graph-derived feature blocks, termed materials (attributes, neighborhood statistics, multi-hop signals, temporal quantities, or directed relations). Section 4 assembles them into a per-node vector $\boldsymbol{m}_i^m$ in Eq. (2). The specification is mechanism-specific, not a label-defined class centroid, and the assembly does not read y, although projections, coordinates, and scores are learned by supervised objectives. Material construction is therefore label-independent, but training is not label-free.

4 METHOD

MSS specifies and organizes graph features according to hypotheses about anomaly mechanisms. One space is one mechanism hypothesis. Its materials are the declared feature blocks, soft occupancy places a node in regions of that space, and a region's typical material is the reference the node is compared against. Together these define a common evidence construction (Sections 4.1–4.4). MSS-T instantiates it as a fixed evidence table consumed by boosted trees. MSS-N learns material projections, private coordinates, and per-space experts, recalibrates regional references, and combines outputs through stop-gradient gated fusion (Section 4.5 and Figure 2).

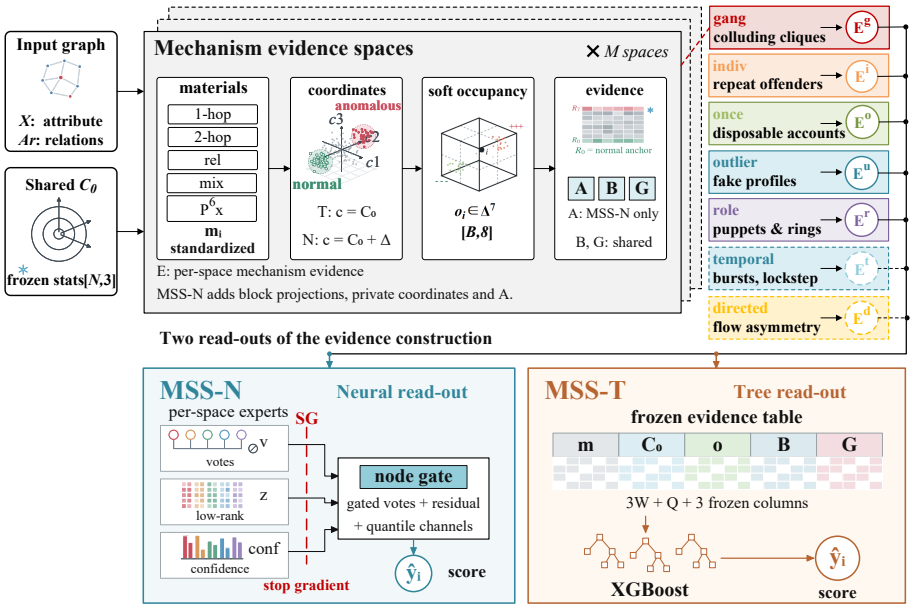

Figure 2: **MSS overview.** Declared materials, soft occupancy, and regional references form per-node evidence, consumed by a tree read-out or a neural one.

4.1 MECHANISM-CONDITIONED SEMANTIC SPACES

The m-th semantic space is an executable mechanism hypothesis $\mathcal{S}_m = (\mathcal{R}_m, \mathcal{B}_m, \mathcal{K}_m)$, where $\mathcal{R}_m$ is an evidence-material specification, $\mathcal{B}_m$ an ordered set of named feature blocks (materials), and $\mathcal{K}_m$ the graph capabilities the hypothesis requires. Each block declares raw or graph-derived features (attributes, multi-hop aggregates, degree and overlap statistics, or topological roles) as an evidence block v and an availability mask a. When time, direction, or a relational capability is absent, the block contributes no evidence instead of a substituted value. Masking, concatenating the n_m blocks, and applying frozen normalization gives

$$\boldsymbol{m}_i^m = \mathrm{Concat}\left(v_i \odot a_i\right) \in \mathbb{R}^{w_m}, \quad (v_i, a_i) = \mathcal{B}_m\big(G, X, i;\, b^{(t-1)}\big), \tag{2}$$

where $b^{(t-1)}$ is the hard bin assignment frozen at the previous recalibration (the mixed-block loop of Figure 2). The tree read-out consumes the declared statistics as they are. The neural read-out first maps each block to a learned projection of width h (Section 4.5).

The specification fixes which materials count as evidence; the construction rules fix how they are organized and compared. The default library contains five capability-universal spaces: gang (collusion), indiv (repeat offending), once (disposable accounts), outlier (fabricated profiles), and role (controllers, puppets, or rings). Temporal and directed spaces are enabled only where the graph declares them. These spaces are hypotheses about anomaly mechanisms, not universal anomaly classes. The gang space aggregates multi-hop means, spread, self blends, and relation shares. The indiv space uses low- and high-pass profiles. The once library declares degree gaps and target reuse. The outlier space uses attribute deviations, and the role space uses role-topology counts (capped triangles, clustering, and k-core). All are frozen, label-free statistics of the permitted context.

The specification is fixed across the reported runs, and its materials, occupancies, references, and deviations remain available with per-space positions and votes in MSS-N.

4.2 Node positions and soft occupancy

Each material channel is standardized with frozen full-graph statistics (Appendix C), which keeps the origin of an axis at population typicality and preserves tail ordering under blocked forward passes. We omit the space superscript. A shared topology backbone C_0, a frozen whitened summary of multi-hop degree statistics (Appendix E), gives a common origin that carries no semantics. The tree read-out uses $c_i = (C_0)_i$, and the neural read-out adds a private learned correction (Eq. (7)). The backbone is computed without labels or community assignments, and the private MSS-N corrections are learned under training-label supervision.

We use three coordinate axes and $Q = 2^3 = 8$ soft bins. Let μ_b and σ_b be frozen full-graph coordinate statistics and $\beta_b(q) \in \{0, 1\}$ the direction of bin q on axis b. The soft occupancy is

$$\begin{aligned} u_{ib} &= \sigma\left(\frac{c_{ib} - \mu_b}{\tau \max(\sigma_b, \varepsilon)}\right), \\ \hat{o}_{iq} &= \prod_{b=1}^{3} u_{ib}^{\beta_b(q)} (1 - u_{ib})^{1-\beta_b(q)}, \qquad o_{iq} = \frac{\hat{o}_{iq}}{\sum_{r=0}^{Q-1} \hat{o}_{ir}}. \end{aligned} \tag{3}$$

Here $\tau = 1$ is the occupancy temperature and $\varepsilon = 10^{-6}$ the numerical floor; $o_i \in \Delta^{Q-1}$ is a differentiable occupancy vector. MSS-N trains toward bin $0 = (0, 0, 0)$ and bin $Q - 1 = (1, 1, 1)$ as normal and anomaly poles; MSS-T assigns no class meaning to these endpoints. The other six bins carry no prescribed class or automatic abstention. The bin directions are fixed, while MSS-N occupancy changes with supervised coordinates. Soft occupancy and its hard-bin summary can be reported with the reference.

4.3 Occupancy-conditioned regional references and evidence

Given the material matrix and soft occupancies of one space, let $d_q = \sum_j o_{jq}$. The typical material of bin q is the occupancy-conditioned average

$$\mathrm{ref}_q = \begin{cases} d_q^{-1} \sum_j o_{jq} \boldsymbol{m}_j, & d_q \geq \varepsilon, \\ \bar{\boldsymbol{m}} = N^{-1} \sum_j \boldsymbol{m}_j, & d_q < \varepsilon. \end{cases} \tag{4}$$

This reference is a statistic, not a learned parameter or a class-conditioned center. The tree computes it once on the frozen coordinates, and the neural read-out recomputes it in a full-graph pass each epoch (Appendix E). Reading the references back through the occupancy and keeping the deviation gives the shared evidence blocks

$$B_i = \sum_q o_{iq} \mathrm{ref}_q, \qquad G_i = \boldsymbol{m}_i - B_i, \tag{5}$$

where B_i is the typical material of the regions the node occupies and G_i its node–reference deviation. With the material and the position they form a per-node mechanism evidence vector, not a pre-existing weak-label view.

4.4 Regional read-out and complexity

Let $H \in \mathbb{R}^{N \times w}$ stack node materials and $O \in \mathbb{R}^{N \times Q}$ stack soft occupancies. Define the row-stochastic reference matrix

$$T_{qj} = \begin{cases} o_{jq}/d_q, & d_q \geq \varepsilon, \\ 1/N, & d_q < \varepsilon, \end{cases} \qquad R = TH, \qquad B = OR = OTH. \tag{6}$$

Equation (6) forms Q regional references and reads them back through the node occupancy: not edge-wise propagation, and no explicit node-pair matrix. For a fixed number of bins it costs $O(NQw)$ time and $O(Qw)$ auxiliary memory, and each sparse multi-hop material builder adds $O(Ew)$ edge-level work.

Proposition 1 (Read-out structure and degeneracy anchors)**.** *Under frozen statistics, the implicit node relation $S = OT$ in Eq.* (6) *is row-stochastic and satisfies* $\mathrm{rank}(S) \le Q$ *without being materialized. At $\lambda = 0$, the anchored block of Eq.* (8) *exactly recovers the unanchored construction. With zero-initialized fusion output layers, the initial fused score is the uniform expert-vote average.*

Appendix F gives the proof; these degeneracies motivate the ablations and the complexity claims.

4.5 Tree and neural read-outs

Tree read-out (MSS-T). The tree read-out trains no representation. It concatenates the deduplicated materials (W columns), the backbone coordinates, the soft occupancy over C_0, and the column-wise B and G of Eq. (5) into one table of width $3W + Q + 3$, and fits a single boosted-tree classifier with library defaults and class-balanced sample weights. Every column is a frozen, label-free statistic.

Neural read-out (MSS-N). The neural mode shares the material specifications and construction rules, with learned representations. Each block is mapped to a learned projection of width h, so $w_m = n_m h$ in Eq. (2), and a private coordinate head consumes a projection of the node attributes and the assembled material:

$$\begin{aligned} g_i &= [\tanh(W_x x_i) \,|\, \boldsymbol{m}_i], \\ c_i &= (C_0)_i + \tanh(W_2 \tanh(W_1 g_i)) \in \mathbb{R}^3. \end{aligned} \tag{7}$$

With ref_0 the reference of the normal pole and λ the anchor strength,

$$\boldsymbol{d}_i = \boldsymbol{m}_i - \lambda \mathrm{ref}_0, \qquad A_i = W_A \,\mathrm{vec}(\boldsymbol{d}_i \otimes o_i), \qquad \mathrm{wv}_i = [A_i \,|\, B_i \,|\, G_i] \in \mathbb{R}^{w_A + 2w}, \tag{8}$$

where A_i combines the direction and magnitude of the departure from the normal profile with the regional position ($\lambda = 1$ in every reported run, and $\lambda = 0$ exactly recovers the unanchored construction of Proposition 1). The space score adds an evidence MLP to the coordinate polarity (Figure 6 in Appendix F):

$$s_i = \frac{\mathbf{1}^\top c_i}{\sqrt{3}} + f_{\mathrm{expert}}(\mathrm{wv}_i), \qquad v_i = \tanh(s_i). \tag{9}$$

MSS-N trains each space with a private objective and passes each expert output to the fusion module through a stop-gradient, so fusion loss cannot change the space representations and standalone scores stay available for diagnosis while fusion learns to combine them. The fused score is

$$s_i^P = \sum_m \beta_{im} v_i^m + f_{\mathrm{res}}(\mathbf{z}_i) + \kappa \sum_m \beta_{im} \tilde{v}_i^m + \kappa' \sum_m \pi_{im} \tilde{v}_i^m, \tag{10}$$

where $\mathbf{z}_i$ contains low-rank summaries of per-space evidence, β_i is produced from these summaries and label-free confidence indicators, $\tilde{v}_i^m$ is calibrated by a frozen quantile table, and π_i is a fixed label-free prior over spaces. The channel weights κ and κ' start at zero, as do the gate and residual output layers, so the initial fused score is the uniform expert-vote average, and fusion loss updates fusion parameters only (Appendix D). All fusion operations are node-local, so full-graph and blocked forward passes are row-wise identical for fixed parameters and frozen statistics (Proposition 2 in Appendix F). This enables million-node processing while retaining per-space diagnostics.

Supervision boundary and training schedule. Each space uses a private class-weighted classification loss, a coordinate-polarity term, and an anti-collapse regularizer, while fusion uses positive-weighted cross-entropy. Raw material extraction, frozen normalization, and Eq. (6) do not read y, whereas projections, coordinates, and scores are supervised. Training has expert, fusion-following, and fusion-autonomous stages. Statistics are recomputed by a no-gradient full-graph pass and then sliced for node blocks (Appendix E).

5 Experiments

We evaluate MSS in four steps: the shared setup (Section 5.1), the fully supervised comparison and its attribution controls (Sections 5.2 and 5.3), the limited-label comparison (Section 5.4), and

deployment cost and mode choice (Section 5.5). Both read-outs are reported throughout, with robustness, efficiency, and full diagnostics in Appendices J, K, and H. These stages answer two questions: which construction parts drive prediction, and whether retained per-space states add information beyond an aggregate score.

5.1 DATASETS, BASELINES, AND PROTOCOL

Datasets. We evaluate on ten graphs from the GADBench suite (Tang et al., 2023), ordered as in the GADBench tables: Reddit (Kumar et al., 2019; Liu et al., 2022), Weibo (Zhao et al., 2020; Liu et al., 2022), Amazon (McAuley & Leskovec, 2013; Dou et al., 2020), YelpChi (Rayana & Akoglu, 2015; Dou et al., 2020), T-Finance (Tang et al., 2022), Elliptic, Tolokers, Questions, DGraph-Fin, and T-Social. Tolokers and Questions come from a heterophily benchmark (Platonov et al., 2023). Elliptic's official split (Weber et al., 2019) is temporal and probes drift (Appendix J). DGraph-Fin (Huang et al., 2022) and T-Social (Tang et al., 2022) are industrial scale.

Baselines. We compare against sixteen fully supervised and eighteen limited-label detectors spanning GADBench defaults and models released in 2024–2026 (roster in Appendix G); normal-only detectors GGAD, RHO, and GraphNC (Qiao et al., 2024; Ai et al., 2025; Zeng et al., 2026) train without labeled anomalies and are excluded rather than adapted.

Protocol. Every detector in Table 1, MSS included, runs under the same split masks and 24GB-GPU cap, ten runs each on seeds $3407 + 10r$, $r = 0, \ldots, 9$. A failed cell counts as zero in the mean, following GADBench. Appendix G details the budget, seed lists, failure markers, and the published-default comparison. AUPRC is the primary metric under label imbalance; Rec@K accompanies the AUROC tables in Appendix H. Under the standard transductive protocol, structure and features of all nodes may freeze label-free statistics, training labels fit, validation labels select, and test labels stay hidden until evaluation. One default configuration serves all datasets, with five spaces in the primary matrix. Temporal and directed materials are supported by both modes and evaluated where the loader declares those capabilities (Table 3), without per-dataset retuning.

5.2 FULLY SUPERVISED COMPARISON

Table 1 reports the fully supervised comparison with both MSS read-outs as separate rows. Neither MSS nor XGB-Graph, the strongest baseline, has a failed cell, so the headline comparison is unaffected by the failure convention. The tree read-out MSS-T attains the highest mean AUPRC, 2.57 points above the strongest baseline, and leads it on nine of the ten graphs. The neural read-out MSS-N reaches 60.10 and leads MSS-T on Reddit and DGraph-Fin (per-graph margins and paired tests: Table 13).

Table 1: **Fully supervised results.** AUPRC ($\times 100$), ten seeds, GADBench order; OOM/OOT count as zero.

Model	Reddit	Weibo	Amazon	Yelp	T-Fin.	Ellip.	Tolo.	Quest.	DGraph.	T-Social	Ave.
MLP	6.51$_{\pm 0.67}$	84.22$_{\pm 2.59}$	90.07$_{\pm 1.90}$	47.56$_{\pm 2.77}$	71.86$_{\pm 2.20}$	22.79$_{\pm 4.57}$	38.15$_{\pm 1.66}$	15.50$_{\pm 1.20}$	2.70$_{\pm 0.02}$	5.06$_{\pm 0.75}$	38.44
GCN	6.94$_{\pm 1.10}$	94.28$_{\pm 1.25}$	35.75$_{\pm 4.40}$	24.31$_{\pm 0.90}$	73.40$_{\pm 2.42}$	21.62$_{\pm 2.56}$	43.31$_{\pm 1.51}$	13.30$_{\pm 1.51}$	3.99$_{\pm 0.09}$	16.13$_{\pm 7.18}$	33.30
GraphSAGE	5.73$_{\pm 1.05}$	83.68$_{\pm 2.08}$	87.35$_{\pm 2.39}$	52.94$_{\pm 2.87}$	61.26$_{\pm 11.31}$	33.27$_{\pm 6.35}$	44.18$_{\pm 1.25}$	16.13$_{\pm 1.12}$	3.77$_{\pm 0.08}$	16.90$_{\pm 3.96}$	40.52
GAT	5.78$_{\pm 0.79}$	90.69$_{\pm 1.94}$	89.69$_{\pm 1.97}$	41.16$_{\pm 3.80}$	61.06$_{\pm 14.31}$	23.80$_{\pm 3.16}$	45.12$_{\pm 1.86}$	16.20$_{\pm 1.65}$	3.88$_{\pm 0.04}$	59.57$_{\pm 5.45}$	43.70
GIN	6.23$_{\pm 0.91}$	89.41$_{\pm 2.00}$	81.87$_{\pm 3.27}$	38.24$_{\pm 2.18}$	65.86$_{\pm 3.21}$	23.42$_{\pm 4.92}$	38.70$_{\pm 0.79}$	13.76$_{\pm 0.89}$	3.26$_{\pm 0.06}$	10.72$_{\pm 1.61}$	37.15
PNA	5.78$_{\pm 0.97}$	92.87$_{\pm 1.29}$	36.33$_{\pm 9.62}$	31.61$_{\pm 1.77}$	34.24$_{\pm 10.74}$	26.59$_{\pm 4.03}$	41.20$_{\pm 1.15}$	12.36$_{\pm 0.73}$	3.45$_{\pm 0.07}$	27.96$_{\pm 7.61}$	31.24
BernNet	7.86$_{\pm 1.28}$	88.95$_{\pm 1.89}$	89.79$_{\pm 1.65}$	50.43$_{\pm 1.52}$	83.09$_{\pm 2.09}$	22.53$_{\pm 3.95}$	43.46$_{\pm 2.40}$	15.01$_{\pm 1.57}$	3.32$_{\pm 0.23}$	18.84$_{\pm 12.61}$	42.33
AMNet	8.01$_{\pm 1.02}$	89.97$_{\pm 1.49}$	89.36$_{\pm 1.90}$	48.82$_{\pm 1.34}$	76.45$_{\pm 2.61}$	14.48$_{\pm 2.62}$	43.45$_{\pm 1.40}$	14.24$_{\pm 1.21}$	2.83$_{\pm 0.01}$	4.04$_{\pm 0.36}$	39.16
BWGNN	**8.53**$_{\pm 1.56}$	91.88$_{\pm 1.32}$	90.11$_{\pm 1.42}$	55.22$_{\pm 3.30}$	87.24$_{\pm 0.81}$	25.63$_{\pm 3.52}$	48.95$_{\pm 1.83}$	16.98$_{\pm 1.67}$	4.00$_{\pm 0.07}$	57.09$_{\pm 11.24}$	48.56
GHRN	8.05$_{\pm 1.64}$	91.62$_{\pm 1.96}$	90.06$_{\pm 1.60}$	55.29$_{\pm 3.48}$	86.97$_{\pm 1.05}$	27.09$_{\pm 5.64}$	49.04$_{\pm 1.80}$	16.85$_{\pm 1.34}$	4.02$_{\pm 0.06}$	OOM	42.90
RF-Graph	5.88$_{\pm 0.35}$	96.95$_{\pm 0.68}$	91.37$_{\pm 1.90}$	75.98$_{\pm 1.14}$	89.19$_{\pm 0.66}$	77.85$_{\pm 0.30}$	48.23$_{\pm 1.08}$	14.09$_{\pm 1.06}$	2.02$_{\pm 0.01}$	93.82$_{\pm 0.06}$	59.54
XGB-Graph	5.97$_{\pm 0.79}$	97.29$_{\pm 0.62}$	93.23$_{\pm 1.23}$	87.21$_{\pm 0.72}$	89.15$_{\pm 0.73}$	**79.16**$_{\pm 0.01}$	48.14$_{\pm 1.24}$	15.78$_{\pm 1.40}$	3.65$_{\pm 0.07}$	92.43$_{\pm 0.10}$	61.20
Recent detectors (2024–2026)											
ConsisGAD	4.22$_{\pm 0.41}$	90.70$_{\pm 1.39}$	91.70$_{\pm 1.72}$	52.36$_{\pm 2.93}$	89.44$_{\pm 0.49}$	46.40$_{\pm 11.89}$	45.19$_{\pm 2.35}$	13.55$_{\pm 1.64}$	3.04$_{\pm 0.11}$	71.67$_{\pm 2.50}$	50.83
SpaceGNN	6.66$_{\pm 0.83}$	94.51$_{\pm 0.84}$	91.64$_{\pm 1.82}$	68.95$_{\pm 1.16}$	90.50$_{\pm 0.60}$	27.00$_{\pm 4.58}$	47.14$_{\pm 1.38}$	19.96$_{\pm 1.45}$	3.94$_{\pm 0.18}$	OOT	45.03
GPA	7.70$_{\pm 1.29}$	95.98$_{\pm 0.63}$	92.60$_{\pm 1.51}$	75.29$_{\pm 0.65}$	89.42$_{\pm 0.46}$	59.39$_{\pm 9.62}$	50.54$_{\pm 1.13}$	19.34$_{\pm 1.69}$	4.04$_{\pm 0.06}$	92.27$_{\pm 0.59}$	58.66
HSMAD	5.72$_{\pm 1.05}$	96.13$_{\pm 1.14}$	93.02$_{\pm 1.09}$	81.16$_{\pm 1.18}$	OOM	48.46$_{\pm 7.92}$	**58.21**$_{\pm 1.70}$	18.45$_{\pm 1.87}$	OOM	OOM	40.12
Ours											
MSS-N	8.50$_{\pm 1.08}$	97.37$_{\pm 0.67}$	92.10$_{\pm 1.61}$	74.93$_{\pm 1.14}$	89.48$_{\pm 0.91}$	57.98$_{\pm 4.02}$	55.90$_{\pm 1.49}$	20.95$_{\pm 2.35}$	**4.53**$_{\pm 0.12}$	99.27$_{\pm 0.03}$	60.10
MSS-T	6.10$_{\pm 0.90}$	**98.16**$_{\pm 0.54}$	**93.94**$_{\pm 1.37}$	**89.79**$_{\pm 0.67}$	**90.71**$_{\pm 0.55}$	77.95$_{\pm 0.00}$	56.27$_{\pm 1.45}$	**20.98**$_{\pm 1.49}$	4.26$_{\pm 0.09}$	**99.53**$_{\pm 0.02}$	**63.77**

On the low-prevalence graphs, AUPRC and AUROC describe different parts of the ranking. MSS-N attains the highest AUROC on Reddit (71.05) and DGraph-Fin (79.09), and MSS-T on Questions (74.95), each highest in its column of the ten-graph roster (Table 11 in Appendix H). These means supplement the AUPRC comparison instead of replacing it.

5.3 Evidence attribution

Complete systems cannot establish which change produced the 2.57-point margin. Fixing the readout at one CPU tree head isolates it over three feeds: two-hop baseline features, those plus our references, and the full MSS evidence. With baseline features the ten-graph mean AUPRC is 62.25 under full supervision and 43.73 under 20+80 labels. Adding our references raises these to 63.04 and 47.31, and the complete evidence reaches 63.89 and 47.90. References therefore contribute 0.80 and 3.59 points and the full construction 1.65 and 4.17. Complete evidence also improves the Random Forest baseline by 0.94 and 3.01 points. The reference effect is concentrated: T-Social alone supplies 5.65 of the 7.98 summed points under full supervision and 31.52 of the 35.90 under 20+80, where its baseline feed scores 41.12 and references alone reach 72.64. Elsewhere the change stays within two points, and dropping T-Social leaves MSS-T first under both budgets (59.79 against XGB-Graph's 57.73, and 44.95 against RF-Graph's 42.45). Appendix J gives the three-graph local view of this spread on the same seed list.[1] The margin over the strongest baseline is not uniform either: Tolokers, T-Social, and Questions supply 80% of the summed per-graph gain. On T-Social, cohesion (gang) and activity (indiv) each exceed 98 AUPRC as standalone views. Elliptic remains the only negative margin, 1.21 below the strongest baseline, under a temporal split that favors tabular tree models.

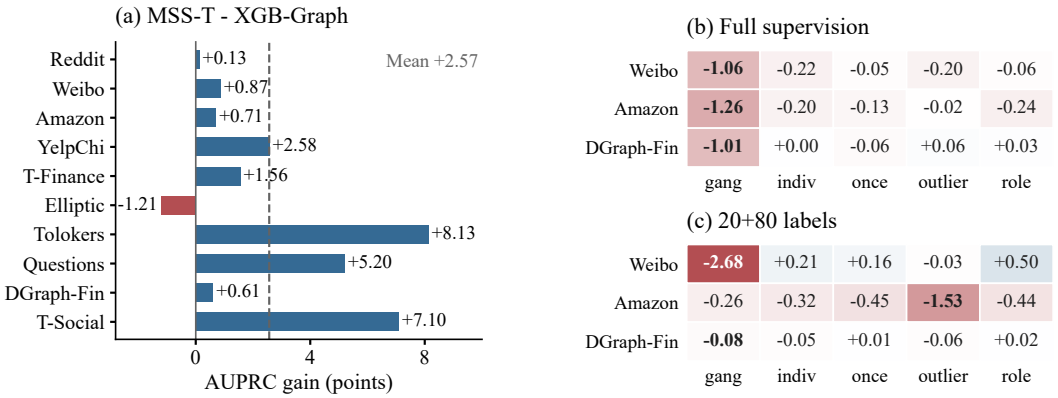

Figure 3: **Attributing the MSS-T margin.** (a) AUPRC gain over XGB-Graph, dashed mean +2.57. (b,c) Material-removal controls; bold marks the largest drop per row.

Figure 3(b,c) shows MSS-T material-removal controls. Removing gang materials costs 1.06, 1.26, and 1.01 AUPRC on fully supervised Weibo, Amazon, and DGraph-Fin. With limited labels the gain comes from Weibo, Amazon, YelpChi, and T-Social, whereas T-Finance, Elliptic, and Questions reverse. The largest losses are gang on Weibo (2.68) and raw attributes on Amazon (1.53). The reference controls (Table 17 in Appendix J) are not uniformly beneficial: removing references reduces Amazon by 0.91 but improves Weibo by 0.40. These controls are interventions, not independent mechanism effects. Appendix J provides all values and the neural removal controls.

5.4 Limited-label evaluation

We reduce training supervision to 20 anomalies and 80 normals without retuning the MSS configuration. MSS-T attains the highest ten-graph mean AUPRC, 47.86, against 44.01 for MSS-N and 43.17 for RF-Graph (Table 2), and reaches a mean AUROC of 80.60. The advantage is not universal: MSS-N leads MSS-T on Weibo and DGraph-Fin, while MSS-T leads on the other eight graphs (Appendix H). A label-budget scan varies only the training subset beyond this 20+80 protocol (Figure 12).

[1] On the seven graphs where every detector completed ten runs, MSS-T still leads (63.31 against 60.97).

Table 2: **Semi-supervised results.** Ten-dataset means under the GADBench 20+80 budget; †APF fails preprocessing on two graphs; Δ: MSS-T minus the best baseline of that row.

	Baselines						Ours		
	RF-Graph	XGB-Graph	ConsisGAD	SpaceGNN	GPA	APF†	**MSS-N**	**MSS-T**	Δ
AUPRC	$43.17_{\pm 0.96}$	$41.99_{\pm 1.07}$	$38.88_{\pm 0.98}$	$42.35_{\pm 1.21}$	$40.18_{\pm 1.04}$	$36.94_{\pm 1.15}$	$\underline{44.01}_{\pm 1.08}$	$\mathbf{47.86}_{\pm 1.47}$	+4.69
AUROC	$78.71_{\pm 0.68}$	$77.55_{\pm 0.87}$	$77.82_{\pm 0.91}$	$\underline{79.14}_{\pm 0.68}$	$78.30_{\pm 0.62}$	$62.87_{\pm 0.72}$	$78.89_{\pm 0.53}$	$\mathbf{80.60}_{\pm 0.85}$	+1.46

5.5 Scalability and mode selection

Industrial-scale graphs. DGraph-Fin (3.70M nodes) and T-Social (5.78M nodes) (Huang et al., 2022; Tang et al., 2022) are an order of magnitude larger than the other eight graphs. MSS attains the best AUPRC and AUROC on both, on DGraph-Fin through the temporal and directed spaces its structure activates (Table 3). By contrast, APF (Liu et al., 2026b) and SAGAD (Liu et al., 2026c) fail the six-hour preprocessing cap, HSMAD the memory cap, and SpaceGNN the 72-hour cap on T-Social.

Table 3: **Industrial graphs.** Main-table summary and temporal/directed capability extensions; ten-run means; failure markers follow the common budget.

	DGraph-Fin (3.70M)			T-Social (5.78M)		
Model	AUPRC	AUROC	Rec@K	AUPRC	AUROC	Rec@K
GCN	$3.99_{\pm 0.09}$	$75.82_{\pm 0.22}$	$7.06_{\pm 0.27}$	$16.13_{\pm 7.18}$	$83.06_{\pm 7.28}$	$21.49_{\pm 10.20}$
BWGNN	$4.00_{\pm 0.07}$	$76.23_{\pm 0.10}$	$7.46_{\pm 0.21}$	$57.09_{\pm 11.24}$	$92.56_{\pm 4.20}$	$56.31_{\pm 7.00}$
TGAT	$4.57_{\pm 0.05}$	$79.75_{\pm 0.11}$	$\underline{8.05}_{\pm 0.20}$	–	–	–
RF-Graph	$2.02_{\pm 0.01}$	$65.59_{\pm 0.21}$	$3.19_{\pm 0.05}$	$93.82_{\pm 0.06}$	$99.16_{\pm 0.01}$	$87.62_{\pm 0.11}$
XGB-Graph	$3.65_{\pm 0.07}$	$74.30_{\pm 0.15}$	$6.77_{\pm 0.32}$	$92.43_{\pm 0.10}$	$99.26_{\pm 0.01}$	$85.28_{\pm 0.14}$
GPA	$4.04_{\pm 0.06}$	$76.49_{\pm 0.29}$	$6.66_{\pm 0.25}$	$92.27_{\pm 0.59}$	$99.02_{\pm 0.12}$	$85.92_{\pm 0.61}$
SpaceGNN	$3.94_{\pm 0.18}$	$75.85_{\pm 0.60}$	$7.55_{\pm 0.42}$		OOT (> 72 h)	
HSMAD		OOM			OOM	
APF, SAGAD		PREP (> 6 h)			PREP (> 6 h)	
Ours						
MSS-N	$4.53_{\pm 0.12}$	$79.09_{\pm 0.18}$	$7.50_{\pm 0.54}$	$\underline{99.27}_{\pm 0.03}$	$\underline{99.94}_{\pm 0.01}$	$\underline{96.65}_{\pm 0.10}$
MSS-N + capability	$\mathbf{4.95}_{\pm 0.07}$	$\mathbf{80.95}_{\pm 0.20}$	$\mathbf{8.19}_{\pm 0.38}$	–	–	–
MSS-T	$4.26_{\pm 0.09}$	$77.55_{\pm 0.20}$	$7.34_{\pm 0.32}$	$\mathbf{99.53}_{\pm 0.02}$	$\mathbf{99.96}_{\pm 0.00}$	$\mathbf{97.42}_{\pm 0.05}$
MSS-T + capability	$\underline{4.67}_{\pm 0.12}$	$\underline{80.06}_{\pm 0.21}$	$7.79_{\pm 0.29}$	–	–	–

On DGraph-Fin, adding the metadata-declared temporal and directed spaces raises MSS-N by 0.42 AUPRC, 1.86 AUROC, and 0.69 Rec@K over its matched base, and MSS-T by 0.36, 2.44, and 0.36 over its own. Both read-outs therefore consume the declared capabilities, and MSS-N attains the higher AUPRC, AUROC, and Rec@K under full supervision (Table 3). Regional aggregation costs $O(NQw)$ (Proposition 1) and evaluation is blocked, avoiding node-pair materialization.

Mode choice and diagnosis. Scenario switches improve mean AUPRC in all six pairs, the largest gain being +3.31 on YelpChi under limited labels, with AUROC trade-offs in Appendix I; merging spaces changes both capacity and objectives (Table 10 in Appendix H). Standalone votes, coordinates, occupancies, scores, and references form the diagnostic views (Figures 7 and 13): the same node follows different paths in different spaces, and the fused probability need not move monotonically, information a merged coordinate system discards. These controls compare space organization jointly with capacity and objectives; the trajectories describe learned representations, not causal motion (Appendix L).

6 Conclusion

MSS declares regional references from anomaly hypotheses and retains view-specific states for diagnosis. Fixed-head controls show that the complete evidence improves two tree read-outs, with gains that depend on the graph and budget. Under 20+80 labels both modes exceed the strongest baseline in ten-graph mean AUPRC. Under full supervision, MSS-T exceeds XGB-Graph on nine graphs, with Elliptic the exception. These results support evidence construction as a design variable alongside the consuming model. MSS remains supervised and transductive, material and reference effects depend on the dataset and the supervision budget, and coordinate motion is a learned diagnostic, not causal confirmation. Improving top-ranked recall while retaining faithful diagnostic states remains an open direction.

REPRODUCIBILITY STATEMENT

All models share one split protocol (a split table exported column by column from the released GAD-Bench dataset files and checksum-verified before every run, using the released budget-specific masks for DGraph-Fin and Elliptic), the common ten-run seed list $3407 + 10r$, $r = 0, \ldots, 9$ (Appendix G), and the declared validation-selection rules (TGAT uses AUROC; neural MSS uses AUPRC); the construction has an exact degeneracy anchor (Proposition 1) and the blocked–full forward equivalence of Proposition 2 is enforced by a runtime check. Diagnostic exports can include the active-space list, per-space scores, coordinates, occupancies, references, fusion weights, and configuration, depending on logging settings. The reported tables are regenerated from saved results; geometric diagnostics require the corresponding states to be retained, and benchmark runs need not save them at every epoch. Configuration files, the per-run results behind every diagnostic table together with the scripts that regenerate the tables from them, and the verification outputs are included as a downloadable archive in the supplementary material; the raw per-run snapshots run to tens of gigabytes and are therefore available to reviewers on request rather than in the archive. Code—including the tuning grid, the verification suite, and a tutorial for writing a new space (the catalogue of available materials, the three-method interface, and a worked example)—is available to reviewers through an anonymized repository during the review period and will be released publicly upon publication. Configuration and the development history of the platform learning rate are disclosed in Appendix G; tree-head sensitivity is reported in Appendix J.

ETHICS STATEMENT

This work studies fraud and abuse detection on public benchmark graphs. No new data were collected and no human subjects were involved; every dataset is used under its published license and carries pseudonymized or already-anonymized node attributes. Anomaly detectors can misclassify legitimate accounts, so the per-node fusion weights and per-space diagnostics that MSS exports are intended to support human review of a flagged account rather than to replace it.

AI USE STATEMENT

Large language models assisted with manuscript drafting and editing, related-work retrieval and checking, code implementation and debugging, experiment scripting and execution under author direction, figure preparation, and checks of numerical results and their interpretation. The authors are responsible for the research design, verification of AI-assisted outputs, interpretation of the evidence, and final content.

REFERENCES

Guoguo Ai, Hezhe Qiao, Hui Yan, and Guansong Pang. Semi-supervised graph anomaly detection via robust homophily learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Leman Akoglu, Mary McGlohon, and Christos Faloutsos. OddBall: Spotting anomalies in weighted graphs. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, 2010.

Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. CopyCatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 2013.

Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning (ICML)*, 2020.

João Bravo. NRGBoost: Energy-based generative boosted trees. In *International Conference on Learning Representations (ICLR)*, 2025.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014.

Ziwei Chai, Siqi You, Yang Yang, Shiliang Pu, Jiarong Xu, Haoyang Cai, and Weihao Jiang. Can abnormality be detected by graph neural networks? In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Nan Chen, Zemin Liu, Bryan Hooi, Bingsheng He, Rizal Fathony, Jun Hu, and Jia Chen. Consistency training with learnable data augmentation for graph anomaly detection with limited supervision. In *International Conference on Learning Representations (ICLR)*, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xiangyu Dong, Xingyi Zhang, Lei Chen, Mingxuan Yuan, and Sibo Wang. SpaceGNN: Multi-space graph neural network for node anomaly detection with extremely limited labels. In *International Conference on Learning Representations (ICLR)*, 2025.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Ben Chamberlain, Michael Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

Yuan Gao, Xiang Wang, Xiangnan He, Zhenguang Liu, Huamin Feng, and Yongdong Zhang. Addressing heterophily in graph anomaly detection: A perspective of graph spectrum. In *Proceedings of the ACM Web Conference (WWW)*, 2023.

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

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In *KI-2012: Poster and Demo Track*, 2012.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. BernNet: Learning arbitrary graph spectral filters via bernstein approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9–10):1641–1650, 2003.

Keith Henderson, Brian Gallagher, Lei Li, Leman Akoglu, Tina Eliassi-Rad, Hanghang Tong, and Christos Faloutsos. It's who you know: Graph mining using recursive structural features. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, and Lei Li. RolX: Structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. DGraph: A large-scale financial dataset for graph anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Junha Kang and Seok-Jun Buu. Graph anomaly detection with disentangled prototypical autoencoder for phishing scam detection in cryptocurrency transactions. *IEEE Access*, 12:91075–91088, 2024. doi: 10.1109/ACCESS.2024.3419152.

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

Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H. Chen. ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2023.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *IEEE International Conference on Data Mining (ICDM)*, 2008.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, Lichao Sun, Jundong Li, George H. Chen, Zhihao Jia, and Philip S. Yu. BOND: Benchmarking unsupervised outlier node detection on static attributed graphs. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference (WWW)*, 2021.

Yilin Liu, Hongchao Zhang, Ahmad F. Taha, Taylor T. Johnson, and Meiyi Ma. Modeling spectral energy shifts in spatio-temporal graph anomaly detection. In *International Conference on Machine Learning (ICML)*, 2026a.

Yunhui Liu, Jiashun Cheng, Yiqing Lin, Qizhuo Xie, Jia Li, Fugee Tsung, Hongzhi Yin, Tao Zheng, Jianhua Zhao, and Tieke He. Towards anomaly-aware pre-training and fine-tuning for graph anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2026b.

Yunhui Liu, Qizhuo Xie, Yinfeng Chen, Xudong Jin, Tao Zheng, Bin Chong, and Tieke He. Mitigating homophily disparity in graph anomaly detection: A scalable and adaptive approach. In *Proceedings of the ACM Web Conference (WWW)*, 2026c.

Zhiwei Liu, Yingtong Dou, Philip S. Yu, Yutong Deng, and Hao Peng. Alleviating the inconsistency problem of applying graph neural network to fraud detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, 2013.

Junjun Pan, Yixin Liu, Xin Zheng, Yizhen Zheng, Alan Wee-Chung Liew, Fuyi Li, and Shirui Pan. A label-free heterophily-guided approach for unsupervised graph fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. arXiv:2502.13308.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations (ICLR)*, 2023.

Hezhe Qiao, Qingsong Wen, Xiaoli Li, Ee-Peng Lim, and Guansong Pang. Generative semi-supervised graph anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tian Qiu, Wenda Li, Zunlei Feng, Jie Lei, Tao Wang, Yi Gao, Mingli Song, and Yang Gao. Association-focused path aggregation for graph fraud detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

Leonardo F.R. Ribeiro, Pedro H.P. Saverese, and Daniel R. Figueiredo. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2020.

Fengzhao Shi, Yanan Cao, Yanmin Shang, Yuchen Zhou, Chuan Zhou, and Jia Wu. H2-FDetector: A GNN-based fraud detector with homophilic and heterophilic connections. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pp. 1486–1494, 2022.

Kijung Shin, Tina Eliassi-Rad, and Christos Faloutsos. CoreScope: Graph mining using k-core analysis—patterns, anomalies and algorithms. In *IEEE International Conference on Data Mining (ICDM)*, 2016.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International Conference on Machine Learning (ICML)*, 2022.

Jianheng Tang, Fengrui Hua, Ziqi Gao, Peilin Zhao, and Jia Li. GADBench: Revisiting and benchmarking supervised graph anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

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

Xuhong Wang, Baihong Jin, Ying Du, Ping Cui, Yingshui Tan, and Yupu Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural Computing and Applications*, 33(18):12073–12085, 2021.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning (ICML)*, 2019.

Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. Semi-supervised credit card fraud detection via attribute-driven graph representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Sheng Xiang, Guibin Zhang, Dawei Cheng, and Ying Zhang. Enhancing attribute-driven fraud detection with risk-aware graph representation. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2501–2512, 2025. doi: 10.1109/TKDE.2025.3543887.

Siyue Xie, Da Sun Handason Tam, and Wing Cheong Lau. CRoC: Context refactoring contrast for graph anomaly detection with limited supervision. In *European Conference on Artificial Intelligence (ECAI)*, 2025.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

Fan Xu, Nan Wang, Hao Wu, Xuezhi Wen, Xibin Zhao, and Hai Wan. Revisiting graph-based fraud detection in sight of heterophily and spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Yiming Xu, Zihan Chen, Zhen Peng, Song Wang, Bin Shi, Bo Dong, and Chao Shen. Generalist graph anomaly detection via prototype-based distillation. In *International Conference on Machine Learning (ICML)*, 2026.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Caiyang Yu, Wei Ju, Haixin Wang, Yifan Wang, and Ziyue Qiao. Dynamic multi-sample mixup with gradient exploration for open-set graph anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Guolei Zeng, Hezhe Qiao, Guoguo Ai, Jinsong Guo, and Guansong Pang. Normality calibration in semi-supervised graph anomaly detection. In *International Conference on Machine Learning (ICML)*, 2026.

Yunhe Zhang, Jinyu Cai, Qi Hao, Pengyang Wang, and See-Kiong Ng. Escaping the homophily trap: A threshold-free graph outlier detection framework via clustering-guided edge reweighting. In *International Conference on Learning Representations (ICLR)*, 2026.

Tong Zhao, Chuchen Deng, Kaifeng Yu, Tianwen Jiang, Daheng Wang, and Meng Jiang. Error-bounded graph anomaly loss for GNNs. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Yue Zhao and Maciej K. Hryniewicki. XGBOD: Improving supervised outlier detection with unsupervised representation learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2018.

Lecheng Zheng, John R. Birge, Haiyue Wu, Yifang Zhang, and Jingrui He. Cluster aware graph anomaly detection. In *Proceedings of the ACM Web Conference (WWW)*, pp. 1771–1782, 2025.

Chen Zhu and Yaying Zhang. HSMAD: Heterophily-driven spectral and manifold learning for graph anomaly detection. In *International Conference on Machine Learning (ICML)*, 2026.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

ROADMAP OF THE APPENDIX

Appendices B–K detail the material library, normalization, fusion, optimization order, and computational analysis. Appendix G specifies the protocol and configurations; Appendix H reports the merged-space control and full predictive results. Appendices I–L cover scenario switching, sensitivity, efficiency, and epoch-wise coordinate cases. The design-axis comparison is summarized in Table 4.

A POSITIONING ALONG DESIGN AXES

Table 4 places MSS and the neighboring routes along four design axes. The axes record design choices, not quality claims. The prototype- and cluster-conditioned routes (Xu et al., 2026; Zhang et al., 2026; Zheng et al., 2025) are unsupervised or zero-shot detectors and therefore fall outside the supervised roster of Appendix G; they are listed because they, too, move the comparison reference away from the edge neighborhood, to learned prototypes or soft clusters in an embedding space, whereas MSS uses fixed, label-free occupancy in MSS-T and supervised coordinates in MSS-N, whose reference statistics are frozen within each epoch. Two further 2025 routes are recorded here for the same reason: HUGE (Pan et al., 2025) is a label-free graph fraud detector and therefore also sits outside the supervised roster, and NRGBoost (Bravo, 2025) replaces the discriminative boosting objective that MSS-T inherits with a generative energy objective over tabular features.

Table 4: Positioning along design axes: where each route places the object a node is compared against, and what it leaves inspectable. *Attrib.*: decisions decompose over named domain mechanisms; *Isol. fusion*: multiple detectors combined without gradient mixing. ✓ = yes, × = no.

Method route	Comparison reference	Inspectable evidence	Attrib.	Isol. fusion
General message passing (Kipf & Welling, 2017)	edge-neighborhood aggregation	shared representation, final score	×	×
GAD-specific message passing (Dou et al., 2020)	selected or reweighted neighborhoods	neighborhood weights, risk evidence	×	×
Spectral and heterophily (Tang et al., 2022)	filtered signals, spectral energy	frequency or heterophily response	×	×
Tree ensembles (Tang et al., 2023)	threshold leaves over graph statistics	decision path, leaf score	×	×
SpaceGNN (Dong et al., 2025)	learned geometric manifold	spatial representation, fused score	×	×
One-class and prototypes (Wang et al., 2021; Qiao et al., 2024)	center, energy, or prototype distance	normality deviation, prototype distance	×	×
Prototype- and cluster-conditioned scoring (Xu et al., 2026; Zhang et al., 2026; Zheng et al., 2025)	learned prototypes or soft clusters in an embedding space	prototype deviation, cluster membership	×	×
MSS (ours)	soft occupancy-conditioned regional reference	materials, occupancy, deviation, per-space vote	✓	✓

B COMPLETE SPACE LIBRARY

The base contract is shared by every space, but the material specification is space-specific. Figure 4 summarizes the library on the feature–topology axis; Table 5 records the current cards. "Equal-weight" means that the graph neighborhood aggregate is the ordinary receiver-side mean; it does not mean that the learned material projection is frozen.

The five A-class spaces are enabled by default. Temporal is explicitly declared and fails at construction if the required time capability is absent. Directed is capability-aware and contributes no fabricated evidence when the graph is undirected or has already been symmetrized. A user-defined

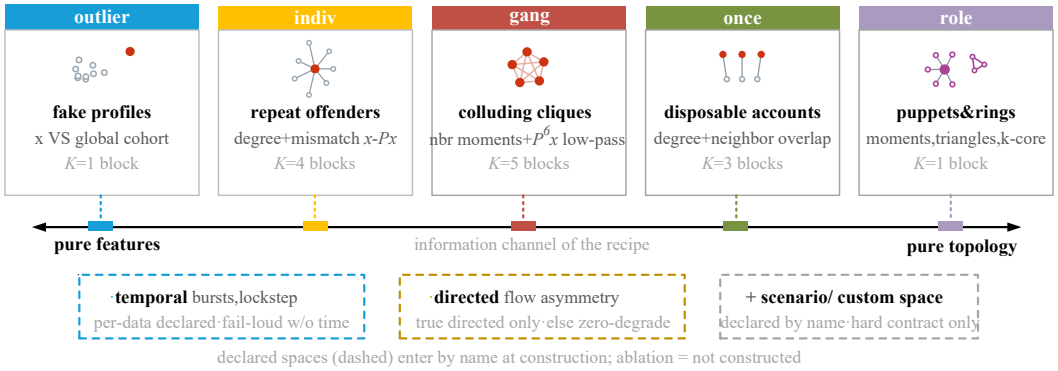

Figure 4: The space library at a glance. Each core space names its anomaly hypothesis (top), its material recipe (middle), and its block count (bottom), and is placed along the axis by how much of its recipe reads features versus topology. Dashed entries are declared spaces: temporal and directed enter by name only where the loader declares the capability (ablation means the space is never constructed), and the scenario slot admits practitioner-authored spaces against the same hard contract (Appendix F).

space follows the same contract: register a subclass, declare named blocks, consume only the read-only view, and enable it at construction. The specifications listed in Table 5 describe the authored material blocks. Every space's coordinate head additionally receives the shared topology anchor and a private three-dimensional projection of the common raw attributes; hence a topology-only specification is not, by itself, a topology-only end-to-end score.

Table 5: Complete space cards. Raw materials are read without labels. Their projections, coordinates, and experts are trained by the corresponding losses.

Space	Anomaly hypothesis	Structural material	Relation/mix/scale material	Blocks	Capability
gang	collusive group	equal-weight neighbor mean/variance, self blend, parallel multi-scale diffusion columns	typed shares, structural-relation coupling, two-hop scale, six-hop low-pass profile P^6x	5	graph
indiv	repeat offender	log degree, self–neighbor mismatch and its population-normalized magnitude, parallel multi-scale diffusion columns	neighbor-degree moments, activity–mismatch coupling, two-hop high-frequency difference	4	graph
once	disposable account	log degree and degree gap	Jaccard overlap and topology-only coupling	3	graph
outlier	fabricated profile	per-column deviation from global cohort (declared classic-scale key)	none	1	attributes
role	controller, puppet, ring	degree moments, capped triangles, clustering, k-core	none	1	topology
temporal	burst or scripted rhythm	node time gradient and neighborhood time variance	edge-time interval and span moments	3	opt. time
directed	harvest, broadcast, pass-through	in/out degree asymmetry and out-degree scale	upstream–downstream feature gap	3	opt. direction

The six-hop low-pass block of the gang space follows a measured single-scale discipline. Adjacent diffusion scales are nearly collinear—the five- and six-hop profiles correlate above 0.99—so exactly one representative scale enters as a named block, and high-pass complements are excluded after they measurably degraded performance on the diffusion-informative dataset. The block enters the full material–occupancy–reference chain rather than being added as a residual, which is what distinguishes it from an additive variant that failed to realize end-to-end gains.

With the default widths (h=32, w_A=32), the weak-view widths are 352 (gang, five blocks), 288 (indiv), 224 (once), 96 (outlier), and 96 (role); enabling temporal or directed extends the set and disabling a space shrinks it automatically.

Material provenance. The materials were derived from the mechanism hypotheses of Section 4.1, not taken from an existing detector; what has precedents in the literature is the *family* each statistic belongs to, not the construction MSS uses. For each material family, Table 6 gives the definition

MSS uses, the space that declares it and the mechanism it evidences, what in the construction is specific to MSS, and the closest family in prior work. The claimed contribution is the specification: which statistic counts as evidence for which mechanism, under which graph capability, with the meaning fixed before evaluation.

Table 6: Material provenance. Each row is one material family: the definition MSS uses (in parentheses), the declaring space and the mechanism it evidences, what in the construction is specific to MSS, and the closest family in prior work (same family, not the same construction). P is the row-normalized adjacency, so $P^k x$ is the receiver-side k-hop mean of x.

Material (definition)	Space · evidence for	MSS-specific construction	Closest family in prior work
Attributes x (raw node attributes)	outlier · fabricated profile	per-column deviation from the global cohort under a declared classic-scale key; no attribute is scored on its own	attribute outlier scores (Goldstein & Dengel, 2012; Li et al., 2023; Liu et al., 2008)
Low-pass means $P^k x$ (k-hop neighborhood means)	gang, indiv · collusive consistency; offender context	scale set $\{1, 2, 3, 6\}$ chosen per mechanism; the six-hop profile enters as one block under a single-scale discipline	precomputed propagation (Wu et al., 2019; Frasca et al., 2020); neighbor aggregates of RF-/XGB-Graph (Tang et al., 2023)
Neighborhood spread (second moment of neighbor attributes)	gang · group cohesion	declared as a cohesion material next to the mean	recursive aggregates (Henderson et al., 2011)
High-pass mismatch $x - P^k x$	indiv · repeat offender	three scales, a population-normalized magnitude, and a two-hop high-frequency difference, declared as materials rather than learned by a filter	high-pass and ego–neighbor separation (Zhu et al., 2020; Tang et al., 2022); local versus aggregate features (Weber et al., 2019)
Self blend $\frac{1}{2}(x + P^1 x)$	gang · collusion	fixed half-weight blend; no learned mixing weight	teleport mixing (Gasteiger et al., 2019); self–neighbor concatenation (Hamilton et al., 2017)
Degree statistics (log degree, mean neighbor degree, degree gap, neighbor-degree moments)	indiv, once · disposable account; activity–mismatch coupling	degree gap defined as mean neighbor degree minus own degree; an activity–mismatch coupling in the indiv space	egonet and recursive degree features (Akoglu et al., 2010; Henderson et al., 2011)
Target reuse and overlap (repeated-target share; Jaccard overlap)	once · lockstep or disposable behavior	repeated-target share as a named material; Jaccard overlap paired with a topology-only coupling	lockstep detection (Beutel et al., 2013; Cao et al., 2014); Jaccard link scores (Liben-Nowell & Kleinberg, 2007)
Role topology (degree moments, capped triangles, clustering, k-core)	role · controller, puppet, ring	six fixed quantities with capped triangle counts; no learned role embedding	role and k-core features (Henderson et al., 2012; Ribeiro et al., 2017; Shin et al., 2016)
Relation composition (typed-relation shares)	gang · multi-relation collusion	shares as a material with a structural–relation coupling, not relation-wise message passing	multi-relation fraud GNNs (Dou et al., 2020; Liu et al., 2021)
Temporal profile (time gradient, neighborhood time variance, edge-time interval and span moments)	temporal, declared · burst or scripted rhythm	three blocks gated by the time capability; no temporal encoder	temporal encoders and burst analysis (Xu et al., 2020; Huang et al., 2022; Kleinberg, 2003)
Directed profile (in/out asymmetry, out-degree scale, upstream–downstream gap)	directed, declared · harvest, broadcast, pass-through	upstream–downstream feature gap; gated by the direction capability	release analysis of DGraph (Huang et al., 2022)
Reference chain $C_0 \to o \to \mathrm{ref} \to (B, G)$	all spaces · organization layer	frozen, parameter-free backbone from whitened multi-hop degree statistics; soft occupancy over $Q = 8$ bins; per-material regional typical value and deviation, refrozen each epoch for MSS-N	cluster- and center-based references (He et al., 2003; Snell et al., 2017; Wang et al., 2021); prototype- and cluster-conditioned scoring (Xu et al., 2026; Zhang et al., 2026; Zheng et al., 2025)

The reported neural harness freezes the explicit once target-reuse projection at zero; the tree feed includes that channel. Shared material specifications therefore do not imply identical consumed columns or learned representations.

C Robust Normalization: Mechanism, Provenance, and Boundary

Material statistics are calibrated once on the graph context permitted by the protocol. To reduce the influence of heavy tails, MSS uses

$$z = \frac{v - \mathrm{med}_{\mathrm{fr}}(v)}{\max\big(1.4826 \cdot \mathrm{mad}_{\mathrm{fr}}(v),\ \varepsilon\big)}, \qquad \mathrm{rz}_{\mathrm{fr}}(v) = \mathrm{sign}(z)\log(1 + |z|), \tag{11}$$

where the median and median absolute deviation are frozen per channel. The constant $1.4826 = 1/\Phi^{-1}(3/4)$ gives Gaussian consistency. Centering places the origin at population typicality; the strictly monotone outer transform compresses tail amplitudes while preserving their order. The log transform is unbounded. The assembled neural material additionally uses a numerical guard: channel d is clipped at $\pm C_d$, where $C_d = \sqrt{\max_i |z_{id}|/\varepsilon}$ is frozen from its calibrated values. The complete pipeline is therefore bounded by this guard. Frozen statistics preserve the blocked-forward semantics of Proposition 2.

A material key may declare *classic-scale*, using frozen mean/std normalization without logarithmic compression. The outlier space declares its deviation key this way because its hypothesis reads cohort-relative amplitude. This key-level rule is identical across datasets; its predictive effect is tested alongside all-classic and no-classic controls in Appendix H.

Development controls exposed two failure modes: uncompressed robust z-scores produced very large material magnitudes, whereas quantile clipping collapsed distinguishable tail values onto one boundary. Log compression addresses these numerical problems but can weaken amplitude-dependent evidence; the classic-scale exception makes that trade-off explicit. Calibration collects channel values at benchmark scale. Industrial graphs use a two-level streaming histogram with $O(D \cdot \mathrm{bins})$ memory. Histogram resolution limits calibration accuracy; fixed, block-additive accumulators preserve the same forward semantics after calibration.

D Fusion Module: Account and Stability

The selected fusion form (Eq. (10)) uses per-space rank-8 projections, a width-16 node gate, a zero-initialized residual head, and calibration scalars κ, κ'. The gate inputs and direct indicator path are specified in Eq. (12) and Figure 5. Vote and indicator quantile tables are non-trainable buffers with $K = 257$ knots: sorted graph values at rounded, uniformly spaced rank positions from 0 to $N - 1$. Piecewise-linear interpolation floors knot spacings at 10^{-12}. They are rebuilt once per epoch and consumed one epoch later. The prior logit scale is fixed at 4. After best-checkpoint restoration, the implementation retains the final-epoch quantile tables; this approximation is part of the reported scoring procedure. Zero initialization gives a uniform vote average. Fusion-side random-state isolation preserves expert trajectories across fusion variants, and parameter counts are obtained from instantiated models.

Written out, the gate and the quantile channels of Eq. (10) are

$$\mathbf{z}_i = \big[\mathrm{wv}_i^1 U_1 \mid \cdots \mid \mathrm{wv}_i^M U_M\big], \qquad \beta_i = \mathrm{softmax}\big(f_{\mathrm{gate}}([\mathbf{z}_i \mid \mathbf{c}_i]) + \mathbf{g}_i W_{\mathrm{out}}\big), \tag{12}$$

with $U_m \in \mathbb{R}^{W_m \times r}$ and $f_{\mathrm{gate}}, f_{\mathrm{res}}$ small multilayer perceptrons. The four indicators per space are endpoint-occupancy decisiveness $o_{i0} + o_{i,Q-1}$ and bounded anchor distance $\tanh(\log(1 + \|\boldsymbol{m}_i - \mathrm{ref}_0\|))$, stacked in $\mathbf{c}_i \in \mathbb{R}^{2M}$, plus vote magnitude $|v_i^m|$ and margin $q_i = \min_b |\sigma((c_{ib} - \mu_b)/\max(\sigma_b, \varepsilon)) - 1/2|$, stacked in $\mathbf{g}_i \in \mathbb{R}^{2M}$. The direct path allows indicator-specific output weights W_{out} to receive gradients from the first step; the indicator inputs remain detached. The calibrated vote is $\tilde{v}_i^m = 2F_m(v_i^m) - 1$, and $\pi_{im} = \mathrm{softmax}_m\{4[F_{|v|,m}(|v_i^m|) + F_{q,m}(q_i^m)]\}$ uses the frozen graph-quantile tables of vote magnitude and margin.

Training schedule. Training follows three stages. Experts train alone during epochs $[0, 30)$. Experts and fusion train together during $[30, 100)$, under a linearly warmed learning rate and a random-state bracket that keeps fusion-side stochasticity from perturbing expert trajectories. Experts then freeze and fusion refines until epoch 200, under cosine decay with validation-AUPRC early stopping at patience 25, counted only in this stage. Expert coordinate statistics and bin references are recomputed once per epoch under no gradient and sliced by training blocks; normalization statistics

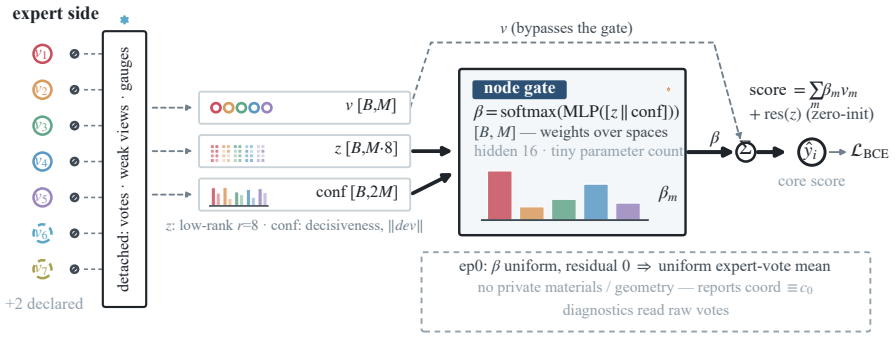

Figure 5: Inside the per-node gate. Experts deliver detached votes $v\,[B, M]$, low-rank weak views $z\,[B, M \cdot 8]$, and label-free gauges $\mathrm{conf}\,[B, 2M]$; the gate MLP (hidden width 16) emits one softmax weight per space per node, and the fused score adds a zero-initialized residual, so at epoch 0 the output is exactly the uniform convex combination of votes (the initialization baseline). The dashed path marks the raw votes. This core schematic omits the direct indicator path and calibrated channels, which are specified in Eqs. (12) and (10).

are calibrated once. During expert stages only, a deterministic stable-pole filter omits a fraction of the longest-resident endpoint nodes from expert reference means, retaining at least eight per pole; fusion, validation, and the frozen-expert stage use unfiltered references (Appendix E).

The two ablated variants are constructed as follows. The anchor-region variant replaces the per-node gate with a learned reliability table $W_{\mathrm{cred}} \in \mathbb{R}^{Q \times M}$ (40 parameters) indexed by the anchor bin of C_0, exporting "which anchor region trusts which space" as a run artifact. The materials-building variant is itself a space. It builds wide three-type materials (attribute, edge-touching, and coupled blocks) with the shared mapping over its own private coordinate, occupancy, and references. The detached votes, scaled by the reliability table and passed through a learned guidance projection, modulate how those materials are assembled, and its objective adds the geometry regularizer of Eq. (15).

Fusion is node-local and excludes batch-statistics normalizers. Expert evaluation and fusion consumption share one block loop, avoiding an $N \times \sum_m W_m$ interface. The contracts and checks in Appendix F cover forward equivalence and gradient isolation.

E Optimization Order and Stable-Pole Filter

Zero-semantic topology anchor. Let d be the receiver degree, $P_{ij} = A_{ij}/\max(d_i, 1)$, and $\mathrm{z_{fr}}$ a z-score whose statistics are frozen on the graph context permitted by the protocol: the observed graph for transductive evaluation and the available past for causal evaluation. The shared anchor of Section 4.2 is

$$\begin{aligned} r_0 &= \mathrm{z_{fr}}(\log(1+d)), \qquad r_1 = Pr_0, \qquad r_2 = Pr_1, \\ C_0 &= \mathrm{stopgrad}\big[\tanh\big(\mathrm{wh_{fr}}([r_0 \,|\, r_1 \,|\, r_2])\big)\big] \in [-1, 1]^{N \times 3}, \end{aligned} \tag{13}$$

where $\mathrm{wh_{fr}}$ denotes frozen whitening: the three stacked columns are mean-centered and multiplied by the inverse Cholesky factor of their jittered covariance $\Sigma + \varepsilon_{\mathrm{jit}} I$, $\varepsilon_{\mathrm{jit}} = 10^{-4}$, decorrelating the axes before the $\tanh$ squashing. An axis whose input variance collapses is zeroed rather than assigned an arbitrary scale. The anchor also conditions the region-gated fusion ablation.

Per-space objective. For a labeled minibatch $\mathcal{L}_B$, let $w_0 = 1$ and $w_1 = n_0/n_1$. The private objective of space m is

$$\mathcal{L}_m = \frac{\sum_{i \in \mathcal{L}_B} w_{y_i} \ell^m_{\mathrm{cls}}(s_i, y_i)}{\sum_{i \in \mathcal{L}_B} w_{y_i}} + \lambda_p \frac{1}{|\mathcal{L}_B|} \sum_{i \in \mathcal{L}_B} \left[\gamma - (2y_i - 1)\frac{\mathbf{1}^\top c_i}{\sqrt{3}}\right]_+ + \mathcal{R}_{\mathrm{geo}}, \tag{14}$$

where the classification loss ℓ^m_{cls} is logistic $\mathrm{softplus}(-(2y-1)s)$ for gang, outlier, role, and temporal; Brier $(\sigma(s) - y)^2$ for once and directed; and confidence-shaped $|\sigma(s) - y|\,\mathrm{softplus}(-(2y-1)s)$ for indiv, with polarity constants $\lambda_p = 0.5$ and $\gamma = 0.5$.

The label-independent anti-collapse regularizer $\mathcal{R}_{\text{geo}}$ of Eq. (14) is

$$\mathcal{R}_{\text{geo}} = \frac{\lambda_f}{Q} \sum_q \left[\frac{\rho}{Q} - \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} o_{iq} \right]_+ + \frac{\lambda_d}{9} \| \operatorname{offdiag}(\operatorname{Corr}(C)) \|_F^2, \tag{15}$$

where $\mathcal{B}$ is the current training block, C stacks its coordinates, and $(\lambda_f, \rho, \lambda_d) = (0.1, 0.25, 0.05)$: a weak per-bin soft occupancy floor (well below uniform, guarding only against severe collapse) and a three-axis decorrelation penalty.

Each epoch runs in a fixed order. **(1)** A no-gradient full-graph or past-prefix pass freezes each expert's coordinate mean, standard deviation, hard bins, and bin references. **(2)** Every unfrozen expert is updated on shuffled labeled blocks with Eq. (14). **(3)** Detached full-graph expert votes are snapshotted for the per-space diagnostic report; in the anchor-region and materials-building variants these snapshots also feed the reliability gating, whose anchor bins are computed once per run from the frozen C_0. **(4)** Only fusion parameters are updated when the schedule permits, inside labeled blocks and under a random-state bracket that keeps fusion-side stochasticity from perturbing expert trajectories. **(5)** Checkpoints are selected by validation AUPRC. Formula-normalization statistics (including the robust median/MAD statistics of every material channel) are calibrated once before this loop.

During expert stages only, a deterministic stable-pole filter is applied to the reference statistics used by expert updates. For node i in space m, the residence counter is

$$t_i^m(e+1) = \begin{cases} t_i^m(e) + 1, & \arg\max_q o_{iq} \in \{0, Q-1\}, \\ 0, & \text{otherwise.} \end{cases}$$

For each pole containing n_p nodes, the $\lfloor c(n_p - k)_+ \rfloor$ longest-resident nodes are omitted from that epoch's pole-reference mean, with deterministic node-index tie breaking, $k = 8$, and $c = 0.25$. At least k nodes remain. The global-mean empty-bin fallback, fusion training, validation, and the frozen-expert stage use the unfiltered references. We report a no-filter ablation because this mechanism changes training references even though it does not change the authored material specification.

F Engineering Contracts and Verification

Every space, including a user-defined one, satisfies six contracts. **(i) Shared coordinate origin**: all spaces start from the same frozen, zero-semantic topology reference. **(ii) Unified output**: the per-node evidence vector $[A \mid B \mid G]$ is the sole material output. **(iii) Material label boundary**: raw extraction, robust normalization statistics, and bin-reference aggregation do not read labels, while learned projections and scores are trained by their losses. **(iv) Read-only views and stop-gradient fusion**: fusion consumes detached votes and detached per-node evidence vectors and never updates expert parameters. **(v) Registry pluggability**: space removal happens at construction, and the fusion consumption follows the active-space set. **(vi) Frozen protocol-level statistics**: normalization (classical and robust) and references are not batch-local and are restricted to past prefixes in the causal protocol. Fusion additionally satisfies a node-local contract: no batch-statistics normalizer and no cross-node operation may enter it. The current common raw-attribute coordinate projection is explicitly outside the authored specification boundary and is tested by a specification-only ablation.

Proposition 2 (Blocked–full forward equivalence)**.** *For fixed parameters and fixed frozen statistics, every forward material quantity ($\boldsymbol{m}_i$, c_i, o_i, ref, $\boldsymbol{d}_i$, A_i, B_i, G_i, and the fusion inputs) is identical whether computed on the full graph or on any partition of nodes into blocks.*

The proof rests on one invariant: after calibration no forward operation aggregates over the batch axis. Graph aggregates are precomputed node-owned quantities, statistics are frozen, and everything else is row-local. The statement concerns forward semantics, not stochastic training trajectories. The transductive protocol may use the observed graph; the strict temporal protocol restricts every statistic and edge to the available past.

Proof of Proposition 1. (i) Each row of O sums to one by the normalization in Eq. (3), and each row of T sums to one by construction (occupancy-normalized rows, uniform fallback rows), so $S = OT$ is row-stochastic; its rank is at most the inner dimension Q. The implementation computes $R = TH$ by a single accumulation over nodes ($O(NQw)$ time, $Q \times w$ auxiliary memory)

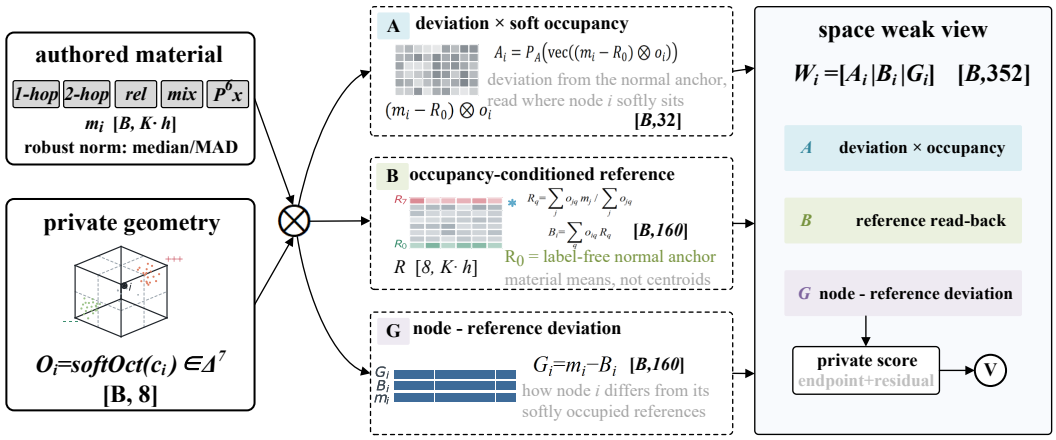

Figure 6: **Per-space evidence assembly.** Authored materials and the private geometry produce the three blocks of contract (ii): the deviation–occupancy interaction A_i, the occupancy-conditioned reference read-back B_i, and the node–reference deviation G_i. Their concatenation $\mathrm{wv}_i = [A_i \,|\, B_i \,|\, G_i]$ is the sole material output, consumed by the private score and, detached, by the fusion gate.

and then $B = OR$ row by row ($O(NQw)$), so S itself is never formed. Graph material builders touch each edge a constant number of times per hop, giving $O(Ew)$ per hop. (ii) With $\lambda = 0$, $\boldsymbol{d}_i = \boldsymbol{m}_i - 0 \cdot \mathrm{ref}_0$; multiplication by zero and subtraction of an exact zero are exact in IEEE arithmetic, so every downstream quantity is numerically identical to the unanchored construction. (iii) The gate and residual heads have zero-initialized output layers, the direct indicator path W_{out} is a zero matrix, and $\kappa = \kappa' = 0$, so at initialization the gate logits vanish, β_i is uniform, $f_{\mathrm{res}} \equiv 0$, both calibration channels contribute zero, and $s_i^P = M^{-1} \sum_m v_i^m$.

Proof of Proposition 2. After calibration, every normalization statistic (robust or classic) is a constant; graph aggregates (neighbor moments, multi-hop diffusion profiles, topology summaries) are precomputed node-owned rows sliced by index; and ref, the coordinate statistics, and the quantile tables of Eq. (10) are frozen within an epoch. Every remaining forward operation (linear maps, elementwise nonlinearities, the outer product $\boldsymbol{d}_i \otimes o_i$, the read-outs $B_i = \sum_q o_{iq} \mathrm{ref}_q$, the indicators, the per-row quantile lookups, and the softmax over the M fusion weights) maps row i's inputs to row i's outputs without reference to any other row in the batch. A composition of row-local maps applied to identical row inputs yields identical row outputs under any partition of rows into blocks.

The verification suite includes a label-permutation probe, permutation equivariance under node reindexing, forward equivalence between full and blocked views (the runtime check of Proposition 2), empty-capability negative tests, expert-parameter non-reachability from the fusion loss, exact degeneracy of the anchored construction at $\lambda = 0$ (Proposition 1(ii)), and intervention-based explanation fidelity. These tests cover implementation claims; they do not replace the predictive and calibration experiments.

G Dataset Cards and Experimental Details

Table 7 anchors the roster: domain, scale, labeled coverage, feature width, positive rate, declared capabilities, and split family for each of the ten graphs.

Baseline roster. The sixteen detectors of Table 1 are: an attribute-only MLP; message-passing GNNs GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2018), GIN (Xu et al., 2019), and PNA (Corso et al., 2020); spectral GAD models BernNet (He et al., 2021), AMNet (Chai et al., 2022), BWGNN (Tang et al., 2022), and GHRN (Gao et al., 2023); the graph-statistics ensembles RF-Graph and XGB-Graph (Tang et al., 2023); and the recent Consis-GAD (Chen et al., 2024), SpaceGNN (Dong et al., 2025), GPA (Qiu et al., 2025), and HSMAD (Zhu & Zhang, 2026).

Table 7: Dataset domain, capability, and loader anchors. Positive rate is over labeled nodes; unknown labels are excluded from its denominator.

Dataset	Domain	Total	Labeled	Feat.	Positive	Rate	Extra cap.	Split
Reddit	social moderation	10,984	10,984	64	366	3.3%	none	stratified
Weibo	synchronized posting	8,405	8,405	400	868	10.3%	none	stratified
Amazon	co-review fraud	11,944	8,639	25	821	9.5%	none	stratified
YelpChi	review fraud	45,954	45,954	32	6,677	14.5%	none	stratified
Tolokers	crowdsourcing	11,758	11,758	10	2,566	21.8%	none	stratified
Questions	Q&A abuse	48,921	48,921	301	1,460	3.0%	none	stratified
T-Finance	finance transactions	39,357	39,357	10	1,803	4.6%	none	stratified
Elliptic	bitcoin transactions	203,769	46,564	166	4,545	9.8%	none	temporal
DGraph-Fin	industrial finance	3,700,550	1,225,601	17	15,509	1.3%	time, dir.	temporal
T-Social	industrial social	5,781,065	5,781,065	10	174,280	3.0%	none	stratified

Origins, following GADBench: Reddit (Kumar et al., 2019; Liu et al., 2022); Weibo (Zhao et al., 2020; Liu et al., 2022); Amazon (McAuley & Leskovec, 2013; Dou et al., 2020); YelpChi (Rayana & Akoglu, 2015; Dou et al., 2020); Tolokers and Questions (Platonov et al., 2023); T-Finance and T-Social (Tang et al., 2022); Elliptic (Weber et al., 2019); DGraph-Fin (Huang et al., 2022).

Compute budget and the three failure markers. Every run is confined to the same budget: one 24GB GPU, twelve pinned logical cores with six BLAS threads, and a 30GB memory cap with swap disabled. The official split masks are verified by checksum before each run, so no detector ever trains on a drifted split. The cap is uniform by construction, so a slower method is never a method that was given a smaller machine. A run that exceeds the memory cap is reported as OOM. For time we project the ten-fold wall clock as the first fold, which carries the one-time graph preprocessing that later folds reuse, plus nine steady-state folds. A projection above 24 hours is flagged, and the run still completes and reports its measured mean; a projection above 72 hours stops once the evidence folds are in and is reported as OOT. Aggregate columns score OOM and OOT as zero. Preprocessing carries its own cap, six hours on the industrial graphs, and a stage that exceeds it stops immediately and is reported as PREP. We report all three rather than omitting the method, because at industrial scale whether a method returns a number is itself a result. The mean aggregate in Table 1 uses one zero convention: all failure markers receive zero scores, so the reported mean rewards complete coverage.

Reruns against the published GADBench defaults. Table 1 reports our own runs rather than the published default-setting means, so the two sets can be compared directly. On the nine graphs other than T-Social both sets use the same default architecture and optimizer settings and differ only in the training budget declared in Table 9, so the twelve methods that appear in both can be read side by side: eight agree within 3.2 points on every one of the nine per-dataset AUPRC values — the two graph-statistics ensembles within 1.1 already inside that band — and four rows exceed it. We list all four rather than the two largest: our GraphSAGE lands 26.5 points above the published default on T-Finance and 18.4 above on Amazon, with a T-Finance standard deviation of 11.3 over the ten seeds, so that configuration is unstable rather than simply better; our BernNet lands 10.7 above on T-Finance and our PNA 6.2 below on Amazon but 5.7 above on T-Finance; our GCN lands 3.2 above on YelpChi, just outside the band. T-Social is listed separately, because that column mixes published defaults with declared adaptations: the two graph-statistics ensembles agree to 0.1 there, GCN, GIN, AMNet, PNA, and BWGNN fall within 2.2, our BernNet lands 13.4 above the published default and our GraphSAGE 4.2 above, our MLP is 9.6 points below — a cell where the published table assigns MLP and GCN the same value while our reruns separate them by 11.1 points — and our own GHRN rerun reports no value there, keeping its memory failure. We keep our own numbers throughout, because mixing the two sources inside one column would prevent a consistent rerun comparison. One result requires a stated adaptation. GAT exceeds the memory cap on T-Social in its default full-graph form, so Table 1 reports its minibatch adaptation (ShaDowKHop sampling, hidden width 8), which completes all ten seeds within the cap and is therefore not a default-setting rerun of the published full-graph configuration. The semi-supervised T-Social cell of GAT in Table 12 uses the same declared adaptation, while DEMO exceeds the budget there even with it. One further entry is capped rather than adapted. SpaceGNN exceeds the 72-hour budget on T-Social with two of ten seeds finished; those two seeds average 98.6 AUPRC, above every completed baseline in that column but below both MSS read-outs. Table 1 scores that cell zero, as it does every OOM entry.

Extended-budget checks on the failure cells. The OOM entries of Table 1 were retried under a wider arm that raises the memory ceiling from 30 to 48 GB and steps the hidden width down to 16, 8, and 4, keeping the split masks and the seed list unchanged. HSMAD completes T-Finance in all ten seeds at 90.67 AUPRC and 97.59 AUROC with width 16; its DGraph-Fin cell also completes, at width 4, but scores 1.36 AUPRC and 50.35 AUROC, indistinguishable from chance; on T-Social the same method exhausts the device at every width, and GHRN repeats its OOM even at width 4. Table 1 keeps the default-configuration entries so that no column mixes two budgets; we report the completed cell here rather than in the table so that the number and the convention that excludes it stay visible together.

Baseline training recipes. Hyperparameters for every baseline follow the authors' public code: whenever the official repository (or the GADBench reference implementation) pins a value, we consume it unchanged, and dataset-specific recipes released by the authors take precedence over global defaults. When a recipe leaves the training length unstated, the detector trains for 200 epochs with early stopping (patience 25, no stopping before epoch 100), the same schedule MSS uses, so no method receives a shorter optimization budget than ours. The 2024–2026 additions (ConsisGAD, Chen et al., 2024; SpaceGNN, Dong et al., 2025; GPA, Qiu et al., 2025; APF, Liu et al., 2026b; SAGAD, Liu et al., 2026c; CRoC, Xie et al., 2025; HSMAD, Zhu & Zhang, 2026; DEMO, Yu et al., 2026; and EGNN, Liu et al., 2026a) run their official recipes verbatim under the split masks and compute budget above, each on the common ten-seed list.

Hyperparameter disclosure. Following the GADBench practice of publishing every consumed configuration, Tables 8 and 9 list the hyperparameters behind every number in the paper. Three properties are worth stating up front. First, MSS uses one shared configuration for all ten datasets and both budgets, and nothing is retuned per dataset. Appendix J scans the tree head around its default. The platform learning rate was fixed at 0.02 during earlier development using a ten-run Reddit stability study that inspected historical test AUPRC; it was then reused unchanged in the reported runs. The earlier validation-based rule had retained 0.005, so the later choice was not preregistered. Changing to the benchmark seed list does not erase this development dependence. Second, the tree read-out consumes the XGBoost library defaults unchanged, so its row cannot hide tuning. Third, baselines inherit the published GADBench default configuration or their authors' released recipes, never a configuration of our choosing.

Table 8: Complete MSS configuration: one shared setting for all ten datasets, both supervision budgets, and both read-outs. No per-dataset retuning is used anywhere in the paper.

Scope	Hyperparameter	Value
shared	semantic spaces M	5 (gang, indiv, once, outlier, role); temporal and directed join only under a metadata declaration
shared	occupancy resolution	$Q = 8$ cells
shared	seeds	$\{3407, 3417, \ldots, 3497\}$ for all model runs.
shared	splits	the released GADBench split columns, exported verbatim from the benchmark's dataset files and checksum-verified before every run; columns 0–9 carry full supervision and 10–19 the 20+80 budget
MSS-N	optimizer	AdamW; one private instance per space plus one for the platform (disjoint moments)
MSS-N	learning rate	5×10^{-3} for spaces; platform peak 2×10^{-2} with linear warmup
MSS-N	schedule	200 epochs: spaces alone $[0, 30)$, spaces + fusion $[30, 100)$, spaces frozen from 100; patience 25, no stop before epoch 100, best-validation checkpoint
MSS-N	batch size	4096 nodes; references rebuilt once per epoch in a no-gradient full-graph pass
MSS-N	dropout	0.3
MSS-N	widths	per-space material block 32, scale block 32, expert head 16; platform $48/96/32$; gate MLP hidden 16; evidence projections $r = 8$
MSS-N	normal anchor	$\lambda = 1.0$
MSS-N	stable-pole filter	floor 8 samples per cell ($= Q$), removal rate $c = 0.25$
MSS-T	boosted head	XGBoost library defaults: depth 6, learning rate 0.3, 100 estimators (the sensitivity scan in Appendix J varies this point)
MSS-T	fitting	histogram method on GPU (XGBoost 1.6.2; refitting the same evidence and splits on a CPU histogram builder under a later XGBoost yields a different ensemble and moves the fully supervised YelpChi ten-run mean by 0.17 AUPRC, so the reported tree rows state their build path), validation on the official split with AUPRC as the metric, class-imbalance sample weight $n_{\text{neg}}/n_{\text{pos}}$ on positives (the ratio of negative to positive training nodes), ten runs

Seeds, splits, and aggregation. The benchmark split table is constructed before training from those released columns. Run r consumes column r under full supervision and $10 + r$ under limited labels; on the eight graphs whose ratios the benchmark resamples those ten columns are distinct

Table 9: Baseline configurations. The GADBench roster runs the published default configuration; the 2024–2026 detectors run their authors' released recipes from vendored official code. Every deviation is declared here or in the rerun paragraph above.

Group	Configuration
GADBench roster (GNNs)	Adam, learning rate 0.01, no dropout, hidden width 32, two layers (AMNet: three blocks; BWGNN/BernNet: order-2 polynomial bases); trained under the uniform budget of 200 epochs, patience 25 after epoch 100, best-validation checkpoint, AUPRC model selection
GADBench roster (trees)	RF-Graph: 100 trees; XGB-Graph: XGBoost defaults (depth 6, $\eta = 0.3$, 100 trees) with GPU histogram; both use the class-imbalance weight $n_{\mathrm{neg}}/n_{\mathrm{pos}}$ on positives, and neither applies early stopping (XGB-Graph only monitors the validation split)
Declared exceptions	T-Social caps hidden width at 16 for every roster GNN (GADBench's own memory guard); GraphSAGE uses the mean aggregator and BernNet two MLP layers (the published table semantics); GAT on T-Social runs the declared ShaDowKHop minibatch adaptation with hidden width 8
Recent detectors (2024–2026)	ConsisGAD, SpaceGNN, GPA, APF, SAGAD, CRoC, HSMAD, DEMO, and EGNN run the architecture, optimizer, schedule, and defaults of their official releases (vendored source); the uniform budget above does not override them. SpaceGNN trains 25 epochs with evaluation at 25, its released CLI default. Declared adaptations: official split masks handed through the data interface, a no-op stub in place of the Weights & Biases logger, and the minibatch retries on T-Social reported with the failure markers
TGAT (DGraph-Fin)	the temporal baseline of the DGraph release, vendored verbatim from the authors' `DGraph_Experiments` code: harmonic time encoding (dimension 32) of relative edge times, one TransformerConv layer (hidden 32, 2 heads, dropout 0.1, edge features = the time encoding), Adam with learning rate 0.01, class-weighted NLL ($1{:}50$), 1000 full-graph epochs with per-epoch validation and the test score taken at the best validation AUROC (the released recipe uses no early stopping); it consumes the same official masks as every other row and runs the ten-seed MSS list
Seeds	the main comparison and controls use $3407 + 10r$, $r = 0, \ldots, 9$, paired by the actual split identity

splits, so a reported deviation spans both split and initialization. DGraph-Fin and Elliptic reuse their fixed partitions under full supervision; their limited-label training, validation, and test masks each vary across the ten runs. The temporal-split description applies to fully supervised Elliptic, not its limited-label masks. All model runs use seeds $3407 + 10r$, $r = 0, \ldots, 9$; changing model seeds does not regenerate masks. Per-graph means are computed from run records; baseline across-graph means retain the benchmark convention of averaging displayed rounded cells, whereas MSS across-graph means use unrounded cells. Standard deviations use divisor n. Aggregate deviations are computed from ten run-indexed dataset means. Ranks use displayed two-decimal means, with ties sharing a rank. Rec@K is recall among the K highest-scoring test nodes, with K equal to the test anomaly count. Neural MSS reports the minimum recall over all orderings of scores tied at the cutoff, recomputed from saved scores. MSS-T, TGAT, and baselines retain archived per-run Rec@K and their original evaluators' tie handling; these values have not been converted to a common tie convention. Nuisance controls retain rounded per-run metrics.

Temporal protocols. Two temporal protocols are kept apart. The standard benchmark is explicitly *transductive*: labels are separated and the observed graph may be used for frozen statistics, and it is not described as forecasting. The strict time-causal protocol rebuilds each evaluation prefix from past nodes, edges, and statistics only. Results from the two protocols are never pooled.

Checkpoints and stored state. Neural MSS checkpoints use validation AUPRC; TGAT follows its released validation-AUROC rule, and tree heads use the fixed estimator count (Table 9). The neural comparisons consistently use exported final-calibration votes from the validation-selected model. These can differ from the trainer's cached validation-best test predictions because export recalibrates the reference state; both are retained where available. Snapshots retain scores, test rows, labels, and configuration. Coordinate histories, occupancies, references, and PDF atlases are optional exports: they require the corresponding logging flags and cannot be reconstructed from final scores alone (Appendix L).

DGraph-Fin protocol boundary. The released fixed split contains $857{,}899$ training, $183{,}862$ validation, and $183{,}840$ test nodes (Huang et al., 2022). Split comparisons require node identities, not just dataset names or counts. We rerun models on the saved masks and common budget rather than treating published scores as a controlled comparison. Capability gains compare enabled and disabled spaces within one protocol and one seed list.

Resource protocol. The tree runner records loading/material construction, reference assembly, and supervised fitting/evaluation separately. Construction is reused within a job; a repeated timer is not ten independent measurements. The merged-space runner records training-stage seconds and

peak CUDA allocated memory, which differs from device-resident memory. Appendix K specifies the measurement boundaries.

H Space Organization and Full Predictive Results

Merged-space MSS-N control. All ten graphs and both budgets use the same original material builders, normalization rules, split masks, seeds, and training schedule. One merged geometry, reference chain, and expert replace the private spaces, with the mean of their original self-loss identities applied to the shared score. This preserves the supplied material groups but is not parameter-matched. The original multi-space mode improves mean AUPRC by 1.38 and 3.38 points; the limited-label exceptions are YelpChi (-0.27) and Tolokers (-0.91). Table 10 reports paired predictions; Table 19 reports the merged mode's resource cost. Its single coordinate system cannot retain independent per-space trajectories, so prediction, diagnostic access, and computation are assessed separately.

Table 10: MSS-N multi-space (M) versus merged-space (S) AUPRC; ten-run means with standard deviations as subscripts.

	Full supervision		Limited labels	
Dataset	M	S	M	S
Reddit	$8.50_{\pm 1.08}$	$7.95_{\pm 0.91}$	$4.58_{\pm 0.48}$	$4.27_{\pm 0.33}$
Weibo	$97.37_{\pm 0.67}$	$96.82_{\pm 0.48}$	$93.26_{\pm 1.62}$	$93.07_{\pm 1.36}$
Amazon	$92.10_{\pm 1.61}$	$91.41_{\pm 1.23}$	$77.50_{\pm 4.24}$	$71.00_{\pm 4.85}$
YelpChi	$74.93_{\pm 1.14}$	$72.57_{\pm 1.45}$	$24.34_{\pm 2.92}$	$24.61_{\pm 2.79}$
T-Finance	$89.48_{\pm 0.91}$	$88.47_{\pm 0.50}$	$74.28_{\pm 2.97}$	$71.42_{\pm 3.52}$
Elliptic	$57.98_{\pm 4.02}$	$52.54_{\pm 6.31}$	$52.34_{\pm 4.64}$	$42.03_{\pm 5.57}$
Tolokers	$55.90_{\pm 1.49}$	$55.89_{\pm 2.10}$	$36.47_{\pm 2.19}$	$37.38_{\pm 2.91}$
Questions	$20.95_{\pm 2.35}$	$17.92_{\pm 2.19}$	$5.66_{\pm 0.93}$	$5.06_{\pm 0.85}$
DGraph-Fin	$4.53_{\pm 0.12}$	$4.42_{\pm 0.06}$	$2.26_{\pm 0.22}$	$1.98_{\pm 0.23}$
T-Social	$99.27_{\pm 0.03}$	$99.24_{\pm 0.05}$	$69.37_{\pm 3.29}$	$55.41_{\pm 10.66}$

Per-space diagnosis before fusion. Figure 7 reports all five space votes on ten graphs and both budgets, each taken from the same fused run. The strongest view changes with the graph and supervision: YelpChi shifts from gang to indiv, and T-Social from indiv to once. Fusion exceeds the largest space mean on every graph under both budgets, but only by 0.01 AUPRC on fully supervised DGraph-Fin. The fully supervised T-Social outlier view trails the strongest space by 46.76 points, revealing a weak view. These scores identify predictive material views; Table 10 tests the effect of merging their organization. Neither comparison establishes that a named space identifies a real-world mechanism.

Mechanism ablations. Figure 8 reports the completed flag controls on the GADBench model-seed list, each paired with a baseline from the same patched code (seven graphs, ten runs, both budgets). The mean control-minus-base AUPRC differences are small and heterogeneous: disabling reliability gating changes Avg_7 by -0.16 under full supervision and $+0.23$ under limited labels; replacing regional references with a global reference gives -0.14 and $+0.17$. All-classic scaling costs 1.67 points on limited-label Weibo but gains 1.53 on T-Finance. These controls support dataset-dependent trade-offs, not universal predictive gains from every component.

Fusion-form controls cover the same seven graphs and ten seeds. Relative to the main-table per-node fusion, the materials-building form has 2.55 and 2.71 lower mean AUPRC under full supervision and limited labels; anchor-region fusion is 0.26 lower and 0.08 higher, respectively. These are complete retraining comparisons. Space removal uses its own matched base and is distinct from merging all materials into one trainable geometry.

Fully supervised track: AUROC and Rec@K. Table 11 completes the fully supervised comparison of Table 1 with the two remaining GADBench metrics, AUROC and Rec@K, computed on the same runs and seeds and under the same zero convention for failure markers. The ordering at

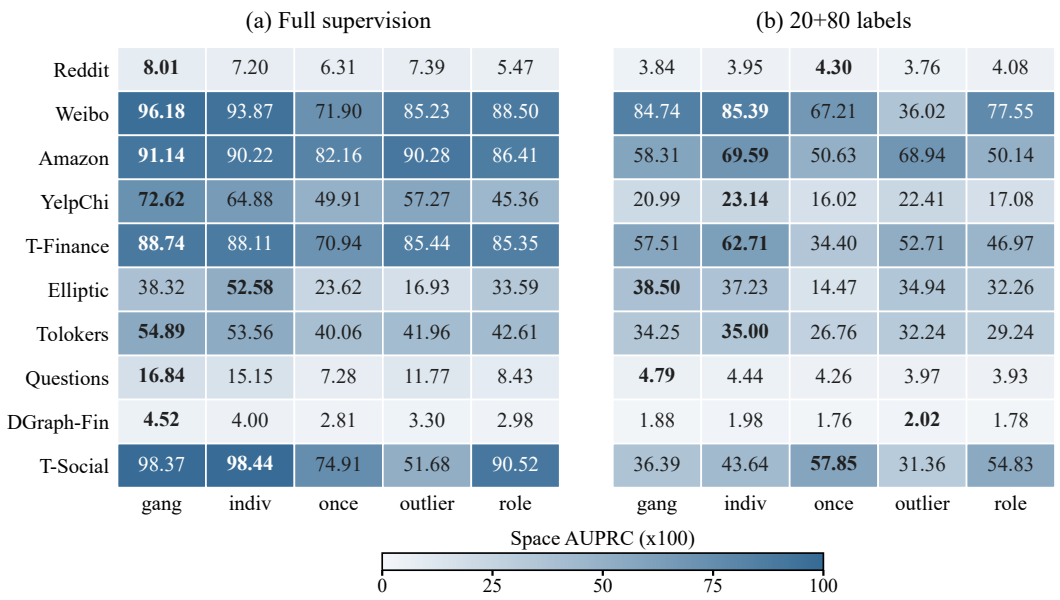

Figure 7: Per-space AUPRC before fusion. Ten graphs, two budgets, ten-run means; bold marks each row's strongest space. Both panels share the 0–100 color scale.

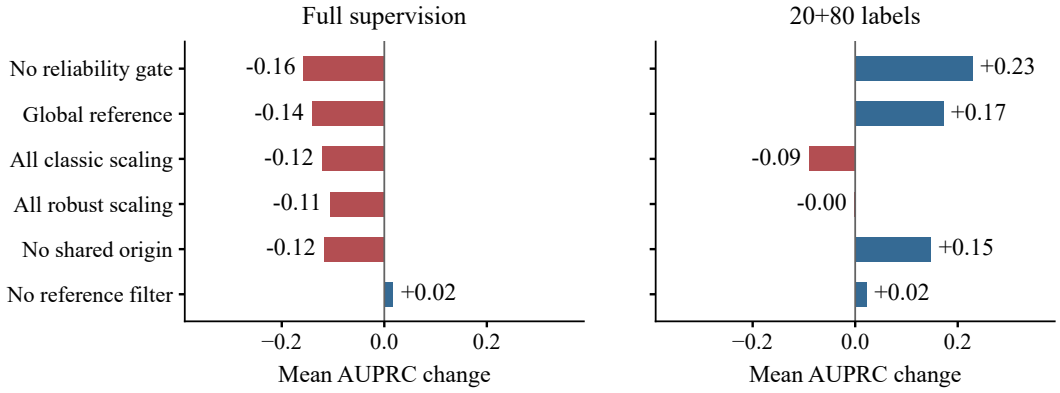

Figure 8: Flag-control ablations on seeds $3407, 3417, \ldots, 3497$. Control minus its matched base, AUPRC ($\times 100$), averaged over seven graphs and ten runs per graph. Positive values favor the control. Both supervision budgets are shown with their corresponding matched bases.

the top does not move: MSS-T holds the best mean on both metrics, leading the strongest baseline by 1.13 points on mean AUROC (GPA) and by 3.17 on mean Rec@K (XGB-Graph). MSS-N has the second-highest mean on both metrics, and the strongest baseline changes with the metric. Beyond the aggregates, the metric swap changes a few per-dataset leaders: MSS-N attains the best Reddit AUROC while the spectral pair BWGNN/GHRN leads Reddit Rec@K, and on Elliptic the tree ensembles stay first on all three metrics. Between the two read-outs, MSS-T leads MSS-N by 19.97 AUPRC on fully supervised Elliptic and 14.86 on YelpChi, while MSS-N leads on Reddit and DGraph-Fin.

Limited-label budget: full tables. Table 12 reports AUPRC, AUROC, and Rec@K for eighteen baselines and both MSS read-outs under the GADBench semi-supervised budget (twenty labeled anomalies plus eighty labeled normals, balanced validation), ten seeds per cell, on all ten graphs. Baselines run the settings their authors recommend: published per-dataset configurations are used verbatim, and unspecified hyperparameters fall back to a uniform budget of 200 epochs, patience 25 after epoch 100, and best-validation checkpointing, with 100 trees for the ensembles. CRoC (Xie et al., 2025) publishes no per-dataset configurations and GPA targets the fully supervised setting;

Table 11: Fully supervised AUROC and Rec@K ($\times 100$), same ten-seed runs as Table 1; best and runner-up per column are bold and underlined. Failure markers count as zero in *Ave.*.

Model	Reddit	Weibo	Amazon	Yelp	T-Fin.	Ellip.	Tolo.	Quest.	DGraph.	T-Social	Ave.
AUROC											
MLP	66.84$_{\pm 2.10}$	90.05$_{\pm 2.19}$	97.87$_{\pm 0.74}$	81.49$_{\pm 1.23}$	91.87$_{\pm 0.79}$	84.27$_{\pm 2.18}$	72.95$_{\pm 1.15}$	69.31$_{\pm 1.29}$	72.44$_{\pm 0.07}$	66.07$_{\pm 2.59}$	79.32
GCN	67.90$_{\pm 3.90}$	98.03$_{\pm 0.65}$	83.25$_{\pm 1.94}$	60.23$_{\pm 1.39}$	91.69$_{\pm 1.51}$	81.36$_{\pm 1.66}$	74.88$_{\pm 0.98}$	70.02$_{\pm 1.24}$	75.82$_{\pm 0.22}$	83.06$_{\pm 7.28}$	78.62
GraphSAGE	65.16$_{\pm 3.78}$	90.00$_{\pm 2.66}$	96.02$_{\pm 1.63}$	83.95$_{\pm 1.12}$	89.10$_{\pm 6.03}$	86.64$_{\pm 1.10}$	77.90$_{\pm 0.58}$	69.44$_{\pm 2.20}$	75.63$_{\pm 0.21}$	83.96$_{\pm 5.09}$	81.78
GAT	64.70$_{\pm 5.47}$	95.08$_{\pm 1.65}$	97.55$_{\pm 0.78}$	78.51$_{\pm 1.74}$	93.13$_{\pm 1.46}$	84.57$_{\pm 1.34}$	78.50$_{\pm 0.95}$	71.27$_{\pm 1.53}$	75.97$_{\pm 0.15}$	95.48$_{\pm 0.32}$	83.48
GIN	68.03$_{\pm 2.23}$	94.67$_{\pm 1.48}$	94.94$_{\pm 0.88}$	77.51$_{\pm 1.21}$	89.30$_{\pm 1.40}$	82.69$_{\pm 2.07}$	74.57$_{\pm 0.54}$	68.98$_{\pm 0.97}$	74.00$_{\pm 0.20}$	78.93$_{\pm 2.75}$	80.36
PNA	65.46$_{\pm 5.33}$	97.23$_{\pm 0.80}$	79.98$_{\pm 6.44}$	74.80$_{\pm 1.38}$	80.17$_{\pm 6.01}$	83.75$_{\pm 1.45}$	74.90$_{\pm 0.54}$	70.99$_{\pm 1.78}$	73.43$_{\pm 0.18}$	81.40$_{\pm 11.58}$	78.21
BernNet	69.03$_{\pm 3.59}$	93.81$_{\pm 1.41}$	97.59$_{\pm 0.58}$	82.75$_{\pm 0.83}$	95.37$_{\pm 0.58}$	83.28$_{\pm 2.00}$	76.67$_{\pm 1.33}$	68.32$_{\pm 2.50}$	74.38$_{\pm 0.57}$	79.14$_{\pm 10.03}$	82.03
AMNet	69.18$_{\pm 3.03}$	95.07$_{\pm 1.13}$	97.72$_{\pm 0.55}$	82.13$_{\pm 0.78}$	94.06$_{\pm 0.74}$	77.19$_{\pm 2.46}$	76.60$_{\pm 0.63}$	68.96$_{\pm 2.52}$	73.07$_{\pm 0.09}$	59.75$_{\pm 3.27}$	79.37
BWGNN	69.31$_{\pm 4.98}$	96.77$_{\pm 0.81}$	98.16$_{\pm 0.40}$	84.63$_{\pm 1.58}$	96.48$_{\pm 0.24}$	85.05$_{\pm 1.24}$	79.72$_{\pm 0.68}$	71.13$_{\pm 2.78}$	76.23$_{\pm 0.10}$	92.56$_{\pm 4.20}$	85.00
GHRN	68.71$_{\pm 4.76}$	96.30$_{\pm 1.14}$	98.24$_{\pm 0.43}$	84.55$_{\pm 1.64}$	96.27$_{\pm 0.23}$	85.46$_{\pm 1.37}$	79.95$_{\pm 0.71}$	70.22$_{\pm 2.76}$	76.13$_{\pm 0.13}$	OOM	75.58
RF-Graph	66.11$_{\pm 2.03}$	99.51$_{\pm 0.16}$	97.59$_{\pm 0.55}$	93.06$_{\pm 0.51}$	96.88$_{\pm 0.38}$	91.91$_{\pm 0.60}$	80.51$_{\pm 0.28}$	65.21$_{\pm 1.29}$	65.59$_{\pm 0.21}$	99.16$_{\pm 0.01}$	85.55
XGB-Graph	64.22$_{\pm 3.71}$	99.57$_{\pm 0.15}$	98.52$_{\pm 0.38}$	96.10$_{\pm 0.25}$	96.71$_{\pm 0.46}$	**93.06**$_{\pm 0.04}$	79.69$_{\pm 0.64}$	65.15$_{\pm 1.39}$	74.30$_{\pm 0.15}$	99.26$_{\pm 0.01}$	86.66
Recent detectors (2024–2026)											
ConsisGAD	56.56$_{\pm 2.18}$	97.10$_{\pm 0.57}$	98.30$_{\pm 0.46}$	83.54$_{\pm 0.94}$	97.00$_{\pm 0.28}$	86.74$_{\pm 1.95}$	78.33$_{\pm 1.02}$	71.14$_{\pm 1.33}$	72.75$_{\pm 0.36}$	96.08$_{\pm 0.18}$	83.75
SpaceGNN	67.59$_{\pm 2.12}$	98.68$_{\pm 0.25}$	98.13$_{\pm 0.42}$	89.84$_{\pm 0.53}$	97.01$_{\pm 0.32}$	84.15$_{\pm 1.49}$	78.79$_{\pm 0.65}$	72.25$_{\pm 0.99}$	75.85$_{\pm 0.60}$	OOT	76.23
GPA	69.42$_{\pm 2.59}$	99.05$_{\pm 0.35}$	98.17$_{\pm 0.62}$	92.19$_{\pm 0.33}$	97.01$_{\pm 0.31}$	89.14$_{\pm 0.93}$	81.23$_{\pm 0.66}$	73.24$_{\pm 1.45}$	76.49$_{\pm 0.29}$	99.02$_{\pm 0.12}$	87.50
HSMAD	59.28$_{\pm 5.42}$	99.06$_{\pm 0.28}$	**98.74**$_{\pm 0.19}$	93.31$_{\pm 0.50}$	OOM	87.34$_{\pm 0.88}$	**84.31**$_{\pm 0.69}$	71.04$_{\pm 4.28}$	OOM	OOM	59.31
Ours											
MSS-N	**71.05**$_{\pm 3.39}$	99.47$_{\pm 0.18}$	98.06$_{\pm 0.46}$	91.66$_{\pm 0.69}$	97.06$_{\pm 0.42}$	87.97$_{\pm 0.62}$	83.17$_{\pm 0.67}$	73.06$_{\pm 2.78}$	**79.09**$_{\pm 0.18}$	99.94$_{\pm 0.01}$	88.05
MSS-T	66.00$_{\pm 3.34}$	**99.61**$_{\pm 0.22}$	98.67$_{\pm 0.46}$	**96.77**$_{\pm 0.21}$	**97.38**$_{\pm 0.33}$	91.88$_{\pm 0.00}$	83.54$_{\pm 0.49}$	**74.95**$_{\pm 1.52}$	77.55$_{\pm 0.20}$	**99.96**$_{\pm 0.00}$	**88.63**
Rec@K											
MLP	8.57$_{\pm 1.88}$	78.67$_{\pm 2.15}$	84.70$_{\pm 1.76}$	46.30$_{\pm 1.94}$	68.23$_{\pm 2.11}$	24.06$_{\pm 11.36}$	38.47$_{\pm 1.90}$	19.71$_{\pm 1.71}$	4.19$_{\pm 0.14}$	3.44$_{\pm 1.82}$	37.63
GCN	10.68$_{\pm 2.92}$	89.16$_{\pm 1.32}$	39.09$_{\pm 4.35}$	25.83$_{\pm 1.14}$	70.18$_{\pm 1.60}$	24.86$_{\pm 5.52}$	42.00$_{\pm 1.56}$	17.19$_{\pm 1.46}$	7.06$_{\pm 0.27}$	21.49$_{\pm 10.20}$	34.75
GraphSAGE	6.73$_{\pm 2.97}$	80.52$_{\pm 2.60}$	83.05$_{\pm 2.96}$	50.21$_{\pm 1.87}$	63.73$_{\pm 9.48}$	40.29$_{\pm 5.48}$	44.29$_{\pm 1.70}$	20.02$_{\pm 1.36}$	7.03$_{\pm 0.26}$	24.25$_{\pm 7.20}$	42.01
GAT	7.21$_{\pm 0.97}$	87.81$_{\pm 1.89}$	84.09$_{\pm 2.08}$	41.72$_{\pm 2.98}$	64.07$_{\pm 9.92}$	27.57$_{\pm 8.37}$	45.06$_{\pm 1.61}$	20.10$_{\pm 2.35}$	7.37$_{\pm 0.26}$	58.70$_{\pm 3.58}$	44.37
GIN	6.33$_{\pm 3.01}$	85.59$_{\pm 1.85}$	76.83$_{\pm 3.14}$	40.41$_{\pm 1.73}$	66.40$_{\pm 2.02}$	27.22$_{\pm 8.34}$	38.04$_{\pm 1.11}$	18.04$_{\pm 1.31}$	5.88$_{\pm 0.30}$	15.53$_{\pm 3.65}$	38.03
PNA	7.62$_{\pm 2.91}$	88.90$_{\pm 1.56}$	41.83$_{\pm 8.67}$	32.17$_{\pm 1.65}$	44.92$_{\pm 13.13}$	31.82$_{\pm 6.24}$	41.81$_{\pm 1.12}$	16.72$_{\pm 1.22}$	6.43$_{\pm 0.26}$	35.65$_{\pm 7.49}$	34.79
BernNet	11.29$_{\pm 1.93}$	84.84$_{\pm 1.73}$	84.21$_{\pm 2.48}$	48.49$_{\pm 1.19}$	78.11$_{\pm 2.04}$	25.32$_{\pm 9.16}$	43.53$_{\pm 1.94}$	19.44$_{\pm 1.90}$	5.71$_{\pm 0.62}$	23.99$_{\pm 13.01}$	42.49
AMNet	12.11$_{\pm 1.42}$	85.36$_{\pm 1.74}$	83.96$_{\pm 3.07}$	47.68$_{\pm 1.26}$	74.99$_{\pm 2.97}$	8.02$_{\pm 5.62}$	42.89$_{\pm 1.13}$	19.88$_{\pm 1.65}$	4.17$_{\pm 0.17}$	4.09$_{\pm 2.03}$	38.31
BWGNN	**13.88**$_{\pm 1.70}$	86.89$_{\pm 2.42}$	84.88$_{\pm 1.30}$	51.75$_{\pm 2.78}$	81.31$_{\pm 1.34}$	31.03$_{\pm 6.48}$	47.85$_{\pm 1.52}$	21.30$_{\pm 1.81}$	7.46$_{\pm 0.21}$	56.31$_{\pm 7.00}$	48.27
GHRN	12.99$_{\pm 1.59}$	86.25$_{\pm 1.87}$	85.30$_{\pm 1.78}$	51.84$_{\pm 2.84}$	81.50$_{\pm 1.13}$	32.95$_{\pm 6.73}$	48.18$_{\pm 1.33}$	21.12$_{\pm 1.53}$	7.45$_{\pm 0.29}$	OOM	42.76
RF-Graph	5.78$_{\pm 1.40}$	92.22$_{\pm 0.96}$	85.30$_{\pm 1.96}$	69.31$_{\pm 1.02}$	84.03$_{\pm 0.82}$	**72.57**$_{\pm 0.32}$	50.43$_{\pm 0.92}$	17.75$_{\pm 1.94}$	3.19$_{\pm 0.05}$	87.62$_{\pm 0.11}$	56.82
XGB-Graph	7.55$_{\pm 1.84}$	92.56$_{\pm 1.11}$	86.22$_{\pm 1.89}$	78.90$_{\pm 0.81}$	84.16$_{\pm 0.78}$	72.47$_{\pm 0.04}$	47.96$_{\pm 1.24}$	17.29$_{\pm 2.39}$	6.77$_{\pm 0.32}$	85.28$_{\pm 0.14}$	57.92
Recent detectors (2024–2026)											
ConsisGAD	4.97$_{\pm 1.67}$	84.41$_{\pm 1.52}$	85.55$_{\pm 1.64}$	49.83$_{\pm 1.23}$	83.90$_{\pm 0.71}$	48.36$_{\pm 12.85}$	45.62$_{\pm 1.95}$	18.41$_{\pm 1.95}$	5.05$_{\pm 0.49}$	65.10$_{\pm 1.99}$	49.12
SpaceGNN	8.30$_{\pm 2.91}$	89.51$_{\pm 1.21}$	84.82$_{\pm 2.15}$	62.28$_{\pm 1.06}$	84.77$_{\pm 0.81}$	37.10$_{\pm 8.51}$	46.47$_{\pm 1.47}$	21.44$_{\pm 2.01}$	**7.55**$_{\pm 0.42}$	OOT	44.22
GPA	10.41$_{\pm 3.90}$	91.18$_{\pm 0.63}$	86.34$_{\pm 2.66}$	67.81$_{\pm 0.89}$	83.92$_{\pm 0.94}$	58.68$_{\pm 6.24}$	49.82$_{\pm 1.43}$	20.81$_{\pm 1.25}$	6.66$_{\pm 0.25}$	85.92$_{\pm 0.61}$	56.16
HSMAD	8.78$_{\pm 3.77}$	92.19$_{\pm 1.14}$	86.46$_{\pm 1.51}$	73.22$_{\pm 0.99}$	OOM	51.63$_{\pm 4.21}$	**56.55**$_{\pm 1.29}$	20.54$_{\pm 1.69}$	OOM	OOM	38.94
Ours											
MSS-N	11.09$_{\pm 2.15}$	92.74$_{\pm 1.43}$	85.06$_{\pm 2.27}$	68.19$_{\pm 0.85}$	84.34$_{\pm 1.23}$	57.27$_{\pm 4.33}$	54.77$_{\pm 1.16}$	23.18$_{\pm 2.80}$	7.50$_{\pm 0.54}$	96.65$_{\pm 0.10}$	58.08
MSS-T	7.48$_{\pm 2.80}$	**93.80**$_{\pm 1.20}$	**87.32**$_{\pm 1.92}$	**81.62**$_{\pm 0.88}$	**85.30**$_{\pm 0.71}$	72.21$_{\pm 0.00}$	54.83$_{\pm 1.09}$	**23.62**$_{\pm 1.88}$	7.34$_{\pm 0.32}$	**97.42**$_{\pm 0.05}$	**61.09**

both run under the same budget as reference rows, as do the two 2026 additions DEMO (Yu et al., 2026) and EGNN (Liu et al., 2026a). MSS-T attains the highest mean AUPRC (47.86) and AUROC (80.60); MSS-N reaches 44.01 and 78.89, respectively.

APF and SAGAD exceed the preprocessing cap on the two industrial graphs, so their ten-dataset means also reflect industrial-scale coverage. DEMO exceeds the memory budget on T-Social even in its declared minibatch adaptation, and a further width-four retry was stopped at the same 24GB ceiling. EGNN crashes on T-Social without returning a result, on its original run and again on a clean serial rerun, so that cell is marked $--$ and scored zero like the other failure markers. The material and reference construction rules are unchanged across budgets. Fewer labels constrain supervised learning, including MSS-N's learned coordinates and experts; label-free aggregation does not make the learned reference structure independent of supervision.

Paired significance tests. The reported margins are means over ten seeds, and both harnesses take trial t from the GADBench seed table under the same split, so the ten runs of a graph are paired and a per-graph margin can be tested directly. Table 13 tests it against the strongest baseline of each ten-dataset mean, XGB-Graph for the fully supervised table and RF-Graph under $20{+}80$, and reports the per-graph margin with the win count and the exact two-sided Wilcoxon signed-rank value; those margins average to the ten-graph mean quoted in the main text. The fully supervised record matches the headline margin on nine of the ten graphs and is significant on eight of them; the one reversal is Elliptic, and there it is significant in the other direction. The limited-label mean is carried by Weibo, Amazon, YelpChi, and T-Social, while T-Finance, Elliptic, and Questions reverse, so the per-graph tests are mixed and a dataset-level sign test does not reject ($p = 0.34$): that claim is about the ten-dataset mean rather than a uniform shift on every graph. Both references are fixed by the ten-graph mean so that a single comparator serves every graph; the best baseline of an individual graph can be a different method, and per-graph best and runner-up are marked in the corresponding tables (Table 1 and Table 12). Read per graph against the best main-table entry on that

Table 12: Semi-supervised AUPRC, AUROC, and Rec@K ($\times 100$), 20+80 labels. Ten-run means with standard deviations as subscripts; best bold, runner-up underlined; failures count as zero in the average.

Model	Reddit	Weibo	Amazon	Yelp	T-Fin.	Ellip.	Tolo.	Quest.	DGraph.	T-Social	Ave.
AUPRC											
MLP	$4.61_{\pm 0.72}$	$59.34_{\pm 3.88}$	$82.26_{\pm 2.45}$	$24.80_{\pm 3.82}$	$50.29_{\pm 13.16}$	$45.40_{\pm 4.92}$	$33.72_{\pm 1.70}$	$6.90_{\pm 1.71}$	$2.23_{\pm 0.22}$	$4.28_{\pm 0.59}$	31.38
GCN	$5.22_{\pm 1.06}$	$89.38_{\pm 2.43}$	$31.98_{\pm 1.71}$	$16.38_{\pm 2.82}$	$60.21_{\pm 9.88}$	$42.58_{\pm 6.80}$	$33.33_{\pm 4.45}$	$5.11_{\pm 1.27}$	$2.17_{\pm 0.32}$	$9.13_{\pm 2.84}$	29.55
GraphSAGE	$4.56_{\pm 0.29}$	$63.53_{\pm 5.77}$	$75.72_{\pm 4.97}$	$24.05_{\pm 4.62}$	$42.06_{\pm 10.08}$	$47.59_{\pm 9.10}$	$35.79_{\pm 1.84}$	$6.28_{\pm 1.29}$	$2.27_{\pm 0.19}$	$6.64_{\pm 1.63}$	30.85
GAT	$4.55_{\pm 0.77}$	$72.61_{\pm 8.49}$	$81.53_{\pm 3.48}$	$25.24_{\pm 3.05}$	$26.57_{\pm 9.83}$	$47.21_{\pm 6.17}$	$34.84_{\pm 1.47}$	$6.83_{\pm 1.34}$	$2.17_{\pm 0.28}$	$10.24_{\pm 2.95}$	31.18
GIN	$4.43_{\pm 0.43}$	$67.40_{\pm 5.71}$	$76.60_{\pm 2.91}$	$24.95_{\pm 3.86}$	$52.46_{\pm 8.20}$	$44.26_{\pm 6.22}$	$34.76_{\pm 1.47}$	$6.17_{\pm 1.12}$	$2.03_{\pm 0.17}$	$6.55_{\pm 1.20}$	31.96
BWGNN	$5.12_{\pm 1.28}$	$77.58_{\pm 5.17}$	$79.91_{\pm 2.95}$	$24.27_{\pm 3.47}$	$69.12_{\pm 10.46}$	$45.79_{\pm 5.62}$	$38.14_{\pm 3.11}$	$6.71_{\pm 1.62}$	$2.10_{\pm 0.11}$	$13.18_{\pm 3.09}$	36.19
AMNet	$\mathbf{5.31}_{\pm 0.62}$	$65.90_{\pm 5.95}$	$81.19_{\pm 1.86}$	$25.25_{\pm 3.05}$	$56.96_{\pm 11.46}$	$30.08_{\pm 4.67}$	$30.40_{\pm 2.34}$	$6.50_{\pm 1.26}$	$2.20_{\pm 0.34}$	$3.29_{\pm 0.32}$	30.71
GHRN	$5.08_{\pm 1.51}$	$75.63_{\pm 7.91}$	$80.46_{\pm 2.05}$	$22.81_{\pm 4.49}$	$66.50_{\pm 11.43}$	$47.67_{\pm 7.05}$	$37.77_{\pm 2.91}$	$6.85_{\pm 1.58}$	$2.18_{\pm 0.16}$	OOM	34.49
RF-Graph	$4.71_{\pm 0.64}$	$75.44_{\pm 9.53}$	$74.56_{\pm 4.35}$	$20.90_{\pm 2.09}$	$80.58_{\pm 2.84}$	$\mathbf{79.07}_{\pm 3.95}$	$37.17_{\pm 2.04}$	$\mathbf{7.50}_{\pm 2.21}$	$2.10_{\pm 0.18}$	$49.70_{\pm 3.74}$	43.17
XGB-Graph	$4.30_{\pm 0.43}$	$73.75_{\pm 6.19}$	$83.25_{\pm 2.42}$	$22.98_{\pm 3.21}$	$76.76_{\pm 3.15}$	$74.90_{\pm 3.75}$	$36.03_{\pm 2.47}$	$6.02_{\pm 1.88}$	$2.03_{\pm 0.25}$	$39.92_{\pm 7.04}$	41.99
ConsisGAD	$4.33_{\pm 0.83}$	$60.60_{\pm 5.08}$	$82.87_{\pm 2.65}$	$25.88_{\pm 1.86}$	$\mathbf{82.74}_{\pm 2.96}$	$46.89_{\pm 7.46}$	$35.37_{\pm 2.45}$	$6.66_{\pm 1.52}$	$2.04_{\pm 0.16}$	$41.46_{\pm 4.51}$	38.88
SpaceGNN	$4.79_{\pm 0.57}$	$80.41_{\pm 2.74}$	$81.04_{\pm 2.36}$	$25.09_{\pm 2.77}$	$81.88_{\pm 2.40}$	$50.25_{\pm 5.43}$	$35.66_{\pm 2.07}$	$6.54_{\pm 1.56}$	$2.07_{\pm 0.27}$	$55.82_{\pm 6.01}$	42.35
CRoC	$3.54_{\pm 0.43}$	$73.26_{\pm 5.31}$	$78.20_{\pm 1.44}$	$22.45_{\pm 2.68}$	$26.47_{\pm 21.52}$	$57.17_{\pm 7.87}$	$34.30_{\pm 2.39}$	$5.73_{\pm 2.05}$	$2.04_{\pm 0.25}$	$3.02_{\pm 0.00}$	30.62
GPA	$4.46_{\pm 0.59}$	$92.73_{\pm 1.22}$	$75.65_{\pm 5.95}$	$28.65_{\pm 3.65}$	$79.02_{\pm 1.40}$	$55.33_{\pm 6.93}$	$37.25_{\pm 2.70}$	$5.96_{\pm 1.67}$	$2.23_{\pm 0.22}$	$20.55_{\pm 4.44}$	40.18
APF	$4.37_{\pm 0.55}$	$91.30_{\pm 2.92}$	$77.97_{\pm 2.76}$	$25.28_{\pm 4.29}$	$78.65_{\pm 3.72}$	$49.64_{\pm 7.67}$	$36.56_{\pm 3.63}$	$5.59_{\pm 0.91}$	PREP	PREP	36.94
SAGAD	$4.97_{\pm 0.75}$	$28.99_{\pm 6.54}$	$79.02_{\pm 2.87}$	$23.44_{\pm 3.79}$	$43.50_{\pm 11.48}$	$50.18_{\pm 7.58}$	$36.84_{\pm 2.87}$	$6.96_{\pm 1.56}$	PREP	PREP	27.39
DEMO	$4.32_{\pm 0.77}$	$68.93_{\pm 7.69}$	$70.82_{\pm 9.95}$	$24.99_{\pm 2.97}$	$53.78_{\pm 7.75}$	$60.68_{\pm 10.91}$	$35.96_{\pm 1.45}$	$6.41_{\pm 1.29}$	$\mathbf{2.33}_{\pm 0.25}$	OOM	32.82
EGNN	$4.01_{\pm 0.51}$	$63.60_{\pm 7.33}$	$45.94_{\pm 8.97}$	$26.63_{\pm 3.27}$	$63.43_{\pm 8.74}$	$42.23_{\pm 6.15}$	$32.88_{\pm 2.26}$	$4.10_{\pm 0.78}$	$1.83_{\pm 0.22}$	--	28.46
MSS-N	$4.58_{\pm 0.48}$	$\mathbf{93.26}_{\pm 1.62}$	$77.50_{\pm 4.24}$	$24.34_{\pm 2.92}$	$74.28_{\pm 2.97}$	$52.34_{\pm 4.64}$	$36.47_{\pm 2.19}$	$5.66_{\pm 0.93}$	$2.26_{\pm 0.22}$	$69.37_{\pm 3.29}$	44.01
MSS-T	$4.77_{\pm 0.78}$	$87.94_{\pm 4.02}$	$\mathbf{84.11}_{\pm 3.10}$	$\mathbf{29.72}_{\pm 5.52}$	$76.55_{\pm 4.66}$	$73.10_{\pm 4.99}$	$\mathbf{40.00}_{\pm 3.73}$	$6.16_{\pm 1.35}$	$2.19_{\pm 0.20}$	$\mathbf{74.09}_{\pm 9.11}$	**47.86**
AUROC											
MLP	$61.39_{\pm 3.37}$	$69.10_{\pm 4.69}$	$91.35_{\pm 2.74}$	$65.30_{\pm 3.95}$	$89.93_{\pm 1.86}$	$90.02_{\pm 0.79}$	$68.14_{\pm 3.02}$	$61.02_{\pm 2.35}$	$\mathbf{68.47}_{\pm 2.00}$	$61.99_{\pm 2.88}$	72.67
GCN	$61.95_{\pm 5.00}$	$96.15_{\pm 1.61}$	$81.16_{\pm 1.42}$	$51.24_{\pm 3.34}$	$88.25_{\pm 2.09}$	$87.18_{\pm 1.42}$	$64.49_{\pm 5.76}$	$58.85_{\pm 4.18}$	$64.38_{\pm 5.51}$	$75.25_{\pm 6.14}$	72.89
GraphSAGE	$62.38_{\pm 1.38}$	$86.46_{\pm 8.17}$	$89.18_{\pm 1.95}$	$64.07_{\pm 5.49}$	$86.58_{\pm 3.02}$	$89.11_{\pm 2.40}$	$71.06_{\pm 1.56}$	$61.86_{\pm 2.54}$	$67.31_{\pm 0.80}$	$71.01_{\pm 5.41}$	74.90
GAT	$61.80_{\pm 3.63}$	$86.45_{\pm 6.90}$	$92.03_{\pm 2.59}$	$65.86_{\pm 3.14}$	$83.17_{\pm 3.72}$	$89.41_{\pm 1.22}$	$69.93_{\pm 1.57}$	$61.95_{\pm 2.50}$	$66.26_{\pm 4.63}$	$73.54_{\pm 4.48}$	75.04
GIN	$62.98_{\pm 2.75}$	$80.67_{\pm 5.77}$	$92.81_{\pm 1.36}$	$64.69_{\pm 4.66}$	$86.55_{\pm 3.24}$	$89.15_{\pm 1.31}$	$71.27_{\pm 1.11}$	$61.98_{\pm 2.41}$	$65.85_{\pm 2.06}$	$71.82_{\pm 2.84}$	74.78
BWGNN	$62.30_{\pm 4.59}$	$93.70_{\pm 2.70}$	$90.94_{\pm 2.33}$	$64.29_{\pm 3.97}$	$92.40_{\pm 1.51}$	$89.36_{\pm 0.97}$	$70.73_{\pm 4.12}$	$63.81_{\pm 4.62}$	$65.70_{\pm 1.56}$	$76.88_{\pm 4.81}$	77.01
AMNet	$\mathbf{63.86}_{\pm 2.32}$	$81.39_{\pm 4.65}$	$91.66_{\pm 2.23}$	$65.90_{\pm 2.77}$	$91.86_{\pm 1.22}$	$84.29_{\pm 1.91}$	$63.72_{\pm 4.62}$	$63.22_{\pm 2.88}$	$66.65_{\pm 3.98}$	$55.00_{\pm 2.97}$	72.75
GHRN	$60.45_{\pm 7.74}$	$92.11_{\pm 6.49}$	$91.42_{\pm 2.22}$	$62.18_{\pm 6.17}$	$92.59_{\pm 0.91}$	$89.52_{\pm 1.33}$	$70.69_{\pm 3.87}$	$63.74_{\pm 4.90}$	$66.66_{\pm 1.67}$	OOM	68.94
RF-Graph	$62.35_{\pm 2.59}$	$96.58_{\pm 1.02}$	$93.13_{\pm 1.49}$	$60.56_{\pm 3.12}$	$\mathbf{94.29}_{\pm 1.46}$	$\mathbf{93.41}_{\pm 2.02}$	$71.94_{\pm 1.70}$	$63.41_{\pm 3.57}$	$64.44_{\pm 2.21}$	$87.01_{\pm 3.61}$	78.71
XGB-Graph	$60.31_{\pm 1.68}$	$95.71_{\pm 0.70}$	$94.24_{\pm 1.54}$	$62.88_{\pm 3.38}$	$93.42_{\pm 1.87}$	$90.73_{\pm 2.44}$	$69.05_{\pm 3.11}$	$60.48_{\pm 3.21}$	$63.14_{\pm 2.92}$	$85.55_{\pm 3.46}$	77.55
ConsisGAD	$58.32_{\pm 4.65}$	$82.76_{\pm 3.84}$	$92.65_{\pm 2.34}$	$66.41_{\pm 2.06}$	$94.16_{\pm 1.39}$	$88.50_{\pm 1.26}$	$70.44_{\pm 1.92}$	$64.67_{\pm 3.29}$	$67.02_{\pm 1.80}$	$93.31_{\pm 1.61}$	77.82
SpaceGNN	$62.86_{\pm 2.44}$	$94.44_{\pm 0.97}$	$92.08_{\pm 1.68}$	$65.53_{\pm 3.09}$	$94.15_{\pm 0.47}$	$89.39_{\pm 1.10}$	$70.55_{\pm 2.65}$	$\mathbf{65.52}_{\pm 2.04}$	$64.03_{\pm 4.16}$	$92.83_{\pm 2.81}$	79.14
CRoC	$53.47_{\pm 4.53}$	$93.34_{\pm 1.86}$	$90.45_{\pm 1.58}$	$60.55_{\pm 3.04}$	$65.34_{\pm 14.70}$	$87.46_{\pm 2.41}$	$68.81_{\pm 3.04}$	$54.91_{\pm 1.98}$	$64.91_{\pm 2.40}$	$50.02_{\pm 0.05}$	68.93
GPA	$60.88_{\pm 2.89}$	$98.35_{\pm 0.45}$	$91.61_{\pm 3.01}$	$66.58_{\pm 2.78}$	$93.71_{\pm 0.92}$	$87.89_{\pm 1.63}$	$\mathbf{72.89}_{\pm 2.26}$	$62.48_{\pm 2.57}$	$65.16_{\pm 2.96}$	$83.46_{\pm 1.60}$	78.30
APF	$60.35_{\pm 3.43}$	$97.78_{\pm 1.07}$	$90.63_{\pm 2.66}$	$65.56_{\pm 4.61}$	$93.09_{\pm 1.37}$	$90.21_{\pm 1.40}$	$71.10_{\pm 3.11}$	$59.96_{\pm 3.81}$	PREP	PREP	62.87
SAGAD	$63.41_{\pm 2.46}$	$73.73_{\pm 6.35}$	$91.86_{\pm 2.30}$	$64.04_{\pm 4.56}$	$87.52_{\pm 3.90}$	$89.13_{\pm 1.94}$	$70.93_{\pm 3.40}$	$62.93_{\pm 3.07}$	PREP	PREP	60.36
DEMO	$61.23_{\pm 4.47}$	$79.65_{\pm 7.87}$	$87.57_{\pm 2.56}$	$65.85_{\pm 2.66}$	$89.09_{\pm 0.86}$	$90.93_{\pm 1.70}$	$71.90_{\pm 1.30}$	$61.79_{\pm 3.33}$	$67.82_{\pm 2.38}$	OOM	67.58
EGNN	$58.58_{\pm 2.52}$	$90.76_{\pm 4.04}$	$81.91_{\pm 7.02}$	$64.51_{\pm 2.96}$	$89.48_{\pm 2.50}$	$85.50_{\pm 3.70}$	$67.12_{\pm 2.00}$	$57.66_{\pm 3.79}$	$59.65_{\pm 3.22}$	--	65.52
MSS-N	$59.86_{\pm 3.17}$	$\mathbf{98.52}_{\pm 0.47}$	$92.58_{\pm 1.34}$	$63.85_{\pm 2.54}$	$92.46_{\pm 0.97}$	$87.22_{\pm 1.42}$	$68.90_{\pm 2.65}$	$63.68_{\pm 2.77}$	$65.63_{\pm 1.98}$	$\mathbf{96.23}_{\pm 0.45}$	78.89
MSS-T	$61.36_{\pm 2.95}$	$97.72_{\pm 0.61}$	$\mathbf{94.95}_{\pm 1.12}$	$\mathbf{70.12}_{\pm 5.41}$	$94.14_{\pm 0.88}$	$89.57_{\pm 2.71}$	$72.61_{\pm 3.41}$	$64.92_{\pm 2.36}$	$64.53_{\pm 2.43}$	$96.09_{\pm 0.48}$	**80.60**
Rec@K											
MLP	$5.61_{\pm 1.74}$	$56.81_{\pm 3.11}$	$\mathbf{78.98}_{\pm 2.16}$	$27.75_{\pm 4.23}$	$58.35_{\pm 10.78}$	$55.07_{\pm 4.88}$	$34.67_{\pm 1.93}$	$10.46_{\pm 3.10}$	$2.67_{\pm 0.99}$	$3.34_{\pm 1.71}$	33.37
GCN	$\mathbf{8.80}_{\pm 2.26}$	$82.10_{\pm 2.84}$	$34.92_{\pm 2.08}$	$16.70_{\pm 3.11}$	$60.91_{\pm 6.52}$	$47.64_{\pm 6.61}$	$34.01_{\pm 4.00}$	$7.46_{\pm 2.50}$	$3.21_{\pm 0.71}$	$10.80_{\pm 6.37}$	30.65
GraphSAGE	$5.12_{\pm 1.80}$	$66.50_{\pm 6.42}$	$72.43_{\pm 5.77}$	$26.81_{\pm 5.55}$	$51.28_{\pm 9.15}$	$52.02_{\pm 7.67}$	$36.86_{\pm 2.05}$	$10.00_{\pm 1.94}$	$3.17_{\pm 0.97}$	$7.61_{\pm 2.95}$	33.18
GAT	$5.18_{\pm 2.33}$	$69.89_{\pm 6.61}$	$78.37_{\pm 2.64}$	$28.23_{\pm 3.41}$	$33.47_{\pm 11.81}$	$53.52_{\pm 4.83}$	$36.00_{\pm 1.57}$	$10.20_{\pm 1.49}$	$2.65_{\pm 0.64}$	$14.21_{\pm 6.79}$	33.17
GIN	$3.87_{\pm 1.40}$	$66.96_{\pm 4.38}$	$72.52_{\pm 3.99}$	$28.11_{\pm 4.53}$	$60.04_{\pm 5.47}$	$50.77_{\pm 5.14}$	$35.81_{\pm 1.40}$	$9.46_{\pm 1.62}$	$2.31_{\pm 0.62}$	$6.99_{\pm 3.28}$	33.68
BWGNN	$7.85_{\pm 2.61}$	$73.68_{\pm 3.53}$	$77.43_{\pm 1.94}$	$27.11_{\pm 3.74}$	$68.64_{\pm 8.56}$	$52.34_{\pm 3.37}$	$38.39_{\pm 2.63}$	$10.89_{\pm 3.12}$	$2.87_{\pm 0.38}$	$21.33_{\pm 4.77}$	38.05
AMNet	$8.07_{\pm 2.27}$	$60.83_{\pm 4.84}$	$77.80_{\pm 2.40}$	$28.48_{\pm 3.19}$	$62.89_{\pm 8.36}$	$33.80_{\pm 8.08}$	$32.06_{\pm 3.07}$	$11.78_{\pm 2.19}$	$2.39_{\pm 0.87}$	$1.59_{\pm 0.58}$	31.97
GHRN	$7.64_{\pm 3.08}$	$72.03_{\pm 4.95}$	$77.86_{\pm 1.97}$	$25.21_{\pm 5.50}$	$68.42_{\pm 6.25}$	$52.86_{\pm 5.01}$	$38.09_{\pm 2.74}$	$11.10_{\pm 2.90}$	$2.88_{\pm 0.60}$	OOM	35.61
RF-Graph	$6.17_{\pm 2.13}$	$66.67_{\pm 7.98}$	$67.21_{\pm 3.86}$	$23.01_{\pm 2.40}$	$74.78_{\pm 2.62}$	$\mathbf{71.35}_{\pm 4.14}$	$39.61_{\pm 1.59}$	$11.55_{\pm 3.05}$	$2.78_{\pm 0.98}$	$48.62_{\pm 2.80}$	41.17
XGB-Graph	$5.15_{\pm 1.71}$	$66.63_{\pm 4.57}$	$78.76_{\pm 2.67}$	$25.65_{\pm 3.99}$	$71.13_{\pm 2.06}$	$67.82_{\pm 3.78}$	$38.07_{\pm 2.43}$	$9.16_{\pm 2.33}$	$3.08_{\pm 0.82}$	$43.51_{\pm 5.85}$	40.90
ConsisGAD	$5.09_{\pm 2.20}$	$54.73_{\pm 6.22}$	$78.95_{\pm 1.13}$	$29.53_{\pm 2.20}$	$\mathbf{79.22}_{\pm 2.51}$	$50.25_{\pm 5.82}$	$36.85_{\pm 2.78}$	$\mathbf{11.82}_{\pm 3.22}$	$1.65_{\pm 0.33}$	$47.75_{\pm 4.32}$	39.58
SpaceGNN	$6.17_{\pm 1.89}$	$72.80_{\pm 2.47}$	$77.27_{\pm 2.33}$	$28.42_{\pm 3.02}$	$78.33_{\pm 2.92}$	$52.44_{\pm 4.62}$	$37.52_{\pm 2.20}$	$10.32_{\pm 2.52}$	$2.82_{\pm 0.53}$	$62.87_{\pm 4.73}$	42.90
CRoC	$3.74_{\pm 0.74}$	$67.14_{\pm 4.40}$	$77.68_{\pm 1.97}$	$25.97_{\pm 2.86}$	$30.30_{\pm 26.30}$	$55.90_{\pm 7.17}$	$35.92_{\pm 2.37}$	$9.59_{\pm 2.75}$	$2.42_{\pm 0.96}$	$2.19_{\pm 0.03}$	31.09
GPA	$5.28_{\pm 1.87}$	$86.28_{\pm 1.48}$	$70.40_{\pm 5.68}$	$30.71_{\pm 3.34}$	$73.98_{\pm 1.97}$	$53.82_{\pm 4.82}$	$37.69_{\pm 2.38}$	$9.18_{\pm 3.15}$	$\mathbf{3.54}_{\pm 0.65}$	$30.06_{\pm 8.48}$	40.09
APF	$6.23_{\pm 2.03}$	$86.03_{\pm 3.87}$	$77.00_{\pm 2.94}$	$27.87_{\pm 4.85}$	$74.65_{\pm 4.64}$	$54.27_{\pm 4.54}$	$37.83_{\pm 3.70}$	$9.08_{\pm 1.78}$	PREP	PREP	37.30
SAGAD	$6.35_{\pm 2.71}$	$37.92_{\pm 8.70}$	$74.64_{\pm 4.61}$	$26.16_{\pm 4.25}$	$51.59_{\pm 8.99}$	$55.85_{\pm 5.26}$	$38.06_{\pm 2.34}$	$11.08_{\pm 2.38}$	PREP	PREP	30.16
DEMO	$5.09_{\pm 2.34}$	$64.59_{\pm 6.56}$	$67.11_{\pm 9.21}$	$27.99_{\pm 3.08}$	$56.91_{\pm 4.13}$	$60.24_{\pm 6.12}$	$37.41_{\pm 1.53}$	$10.09_{\pm 1.94}$	$3.25_{\pm 0.75}$	OOM	33.27
EGNN	$4.51_{\pm 1.55}$	$65.64_{\pm 6.87}$	$52.07_{\pm 7.28}$	$30.49_{\pm 3.31}$	$67.48_{\pm 4.47}$	$47.35_{\pm 5.84}$	$35.14_{\pm 2.57}$	$5.30_{\pm 2.37}$	$2.60_{\pm 0.91}$	--	31.06
MSS-N	$6.44_{\pm 1.20}$	$\mathbf{86.87}_{\pm 3.07}$	$71.22_{\pm 4.98}$	$28.31_{\pm 3.48}$	$70.92_{\pm 2.71}$	$52.65_{\pm 4.02}$	$39.00_{\pm 1.94}$	$9.77_{\pm 2.35}$	$3.45_{\pm 0.55}$	$65.16_{\pm 3.59}$	43.38
MSS-T	$6.47_{\pm 2.21}$	$84.71_{\pm 3.84}$	$78.57_{\pm 2.62}$	$\mathbf{32.31}_{\pm 5.68}$	$71.34_{\pm 2.66}$	$66.08_{\pm 5.09}$	$\mathbf{42.02}_{\pm 3.64}$	$9.68_{\pm 2.59}$	$3.35_{\pm 0.43}$	$\mathbf{72.31}_{\pm 6.62}$	**46.68**

graph, MSS-T leads seven of the ten fully supervised graphs and four of the ten limited-label graphs. The fully supervised shortfalls are Reddit behind BWGNN, Tolokers behind HSMAD, and Elliptic behind XGB-Graph; the limited-label shortfalls are Weibo behind GPA, T-Finance behind ConsisGAD, Elliptic and Questions behind RF-Graph, Reddit behind AMNet by 0.54 AUPRC points, and DGraph-Fin behind DEMO by 0.14 points.

Table 13: Paired per-run tests of MSS-T against the strongest baseline of each ten-dataset mean: XGB-Graph in the fully supervised setting and RF-Graph under $20{+}80$. Trial t of both sides uses seed $3407 + 10t$ under the same split; ΔAUPRC is the difference of the ten-run means as printed in the main tables, w counts the ten runs where MSS-T is higher, and p is the exact two-sided Wilcoxon signed-rank value, bold when $p \leq 0.05$. The last row gives the ten-graph mean margin, the graphs won, and the two-sided sign test over the ten per-graph margins; the fully supervised Elliptic pair uses the archived per-run record of that column.

	Fully supervised vs. XGB-Graph			Limited labels vs. RF-Graph		
Dataset	ΔAUPRC	$w/10$	p	ΔAUPRC	$w/10$	p
Reddit	$+0.13$	6	.43	$+0.06$	4	.85
Weibo	$+0.87$	10	**.002**	$+12.50$	9	**.004**
Amazon	$+0.71$	9	**.027**	$+9.55$	10	**.002**
YelpChi	$+2.58$	10	**.002**	$+8.82$	9	**.004**
T-Finance	$+1.56$	10	**.002**	-4.03	0	**.002**
Elliptic	-1.21	0	**.002**	-5.97	1	**.004**
Tolokers	$+8.13$	10	**.002**	$+2.83$	8	**.037**
Questions	$+5.20$	10	**.002**	-1.34	3	.084
DGraph-Fin	$+0.61$	10	**.002**	$+0.09$	6	.38
T-Social	$+7.10$	10	**.002**	$+24.39$	10	**.002**
Ten-graph mean	$+2.57$	9/10	**.021**	$+4.69$	7/10	.34

Boundary and polarity checks. Label permutation verifies the stated supervision boundary. Per-bin anomaly rates and risk–coverage curves test abstention, and an intermediate region that out-enriches the anomaly endpoint counts as a polarity failure.

Authoring protocol. The frozen authoring specification states the hypothesis, the columns it reads, the graph materials it builds, the capabilities it requires, and the failure domain it accepts, and the authored space is compared against all-column, random-column, missing-block, and specification-only-coordinate controls. We admit only evidence a practitioner can obtain before testing when writing that specification.

I Scenario-Space Switching: Full Paired Results

This ten-run study uses the common GADBench seed list and its own matched scenario baselines. It expands the scenario-space switching summary of Section 5.5. The three spaces are authored from scenario reasoning rather than from label statistics. *Banpattern* on Reddit reads that banned accounts are typically disposable—low-investment profiles, batched one-shot targets, and posts that deviate from the local context. *Coldquestion* on Questions reads that unanswered questions pair cold content with low attention and a context gap to their neighborhood. *Reviewfarm* on YelpChi reads that farmed reviews carry extreme, templated ratings and an unusual relation-composition signature. Table 14 reports the same-seed paired comparison of the five-space base against the six-space configuration with the scenario space switched on, under both supervision protocols, for the neural read-out whose gate consumes the new space directly.

The largest AUPRC gain is reviewfarm on limited-label YelpChi: 3.31 AUPRC, 2.96 AUROC, and 2.44 Rec@K. Coldquestion improves fully supervised Questions by 0.41 AUPRC, but its limited-label AUROC falls by 0.77 despite a 0.12 AUPRC gain. Banpattern improves AUPRC in both budgets, while its limited-label Rec@K falls by 0.18. Thus scenario declarations can help without guaranteeing improvement in every metric. These paired scenario baselines are not replaced by the independently executed main-table row.

Table 14: Scenario switches for MSS-N, matched scenario baselines, ten GADBench-seed runs. AUPRC/AUROC/Rec@K are scaled by 100.

Dataset	Space	Full AUPRC	AUROC	Rec@K	Limited AUPRC	AUROC	Rec@K
Reddit	base	$9.08_{\pm 1.24}$	$72.43_{\pm 2.18}$	$12.18_{\pm 2.81}$	$4.53_{\pm 0.47}$	$59.71_{\pm 3.26}$	$6.38_{\pm 1.10}$
Reddit	banpattern	$9.26_{\pm 1.76}$	$72.38_{\pm 3.54}$	$12.59_{\pm 2.92}$	$4.67_{\pm 0.52}$	$61.17_{\pm 2.92}$	$6.20_{\pm 1.81}$
Questions	base	$20.42_{\pm 2.54}$	$73.12_{\pm 1.45}$	$22.96_{\pm 3.26}$	$5.82_{\pm 0.93}$	$64.14_{\pm 2.49}$	$9.89_{\pm 1.97}$
Questions	coldquestion	$20.83_{\pm 2.37}$	$73.30_{\pm 2.21}$	$23.45_{\pm 2.39}$	$5.94_{\pm 1.21}$	$63.38_{\pm 2.64}$	$10.08_{\pm 2.34}$
YelpChi	base	$75.01_{\pm 0.53}$	$91.83_{\pm 0.44}$	$68.18_{\pm 0.60}$	$24.46_{\pm 3.02}$	$64.02_{\pm 2.61}$	$28.45_{\pm 3.51}$
YelpChi	reviewfarm	$75.43_{\pm 1.05}$	$91.90_{\pm 0.70}$	$68.34_{\pm 0.83}$	$27.77_{\pm 3.14}$	$66.98_{\pm 2.50}$	$30.89_{\pm 3.38}$

Temporal and directed capabilities. Table 15 completes the paired DGraph-Fin capability comparison for both modes and budgets. MSS-T uses the base from its own capability batch rather than subtracting the independent main-table row. Mean AUPRC, AUROC, and Rec@K increase in both modes and budgets, without implying improvement in every run. T-Social carries neither temporal nor directed structure, so neither space can be activated on it, and DGraph-Fin is the only graph in the suite that supports both.

Table 15: Paired DGraph-Fin gains from temporal and directed spaces; ten-run means and standard deviations.

Budget	Mode	ΔAUPRC	ΔAUROC	ΔRec@K
Full	MSS-N	$+0.42_{\pm 0.16}$	$+1.86_{\pm 0.29}$	$+0.69_{\pm 0.74}$
Full	MSS-T	$+0.36_{\pm 0.18}$	$+2.44_{\pm 0.27}$	$+0.36_{\pm 0.52}$
Limited	MSS-N	$+0.11_{\pm 0.18}$	$+1.34_{\pm 1.79}$	$+0.20_{\pm 0.46}$
Limited	MSS-T	$+0.12_{\pm 0.10}$	$+1.59_{\pm 1.14}$	$+0.17_{\pm 0.36}$

J Robustness and Sensitivity Analyses

All controls use the GADBench model-seed list and the stated paired masks.

Temporal drift on Elliptic. Elliptic uses the official temporal split: training steps up to 25, validation steps 26–34, and test steps after 34. Every detector in Table 1 uses this split. MSS-T reaches 77.95 AUPRC, 1.21 points below XGB-Graph; its aggregate advantage therefore does not imply uniform superiority under temporal distribution shift.

Fixed read-outs and evidence construction. Table 16 expands the main-text attribution into ten-graph results, holding the read-out and seeds fixed within each tree family. Regional references augment baseline features; complete MSS also changes the material construction. Both families support mean gains from complete evidence, but the dataset-level effects are heterogeneous. The intermediate RF arm uses MSS materials only; it is distinct from XGBoost's baseline-plus-reference arm. These CPU controls are separate from the GPU main table.

MSS-T materials and construction. The study covers Weibo, Amazon, and DGraph-Fin under both budgets; material removals appear in Figure 3(b,c), and reference controls in Table 17. The five removal arms drop all columns declared by the selected group from the deduplicated table, including columns shared with other groups, and rebuild references. No-reference retains only materials; global-reference replaces the topology backbone by a constant before constructing occupancies and references; neighbor-reference replaces regional references with neighbor averages. Thus the constant-backbone arm changes several linked blocks and does not isolate the backbone alone. The effects in Figure 3(b,c) are conditional on these interventions, not additive mechanism contributions.

Secondary hyperparameter sensitivity. The ten-seed sweep covers depths $\{4, 6, 8\}$, learning rates $\{0.05, 0.1, 0.3\}$, and $\{100, 300\}$ trees on the eight graphs below industrial scale; the two industrial graphs (DGraph-Fin, T-Social) follow the separate protocol of Section 5.5. The sweep

Table 16: Fixed-read-out AUPRC, ten-run means and standard deviations. Base: two-hop features; +Ref: baseline with references; Mat.: MSS materials only; MSS: complete evidence.

	XGBoost			Random Forest		
Dataset	Base	+Ref	MSS	Base	Mat.	MSS
Full supervision						
Reddit	$5.88_{\pm 0.72}$	$6.30_{\pm 0.64}$	$6.18_{\pm 0.75}$	$5.89_{\pm 0.34}$	$6.04_{\pm 0.47}$	$6.20_{\pm 0.49}$
Weibo	$98.01_{\pm 0.54}$	$98.02_{\pm 0.54}$	$98.16_{\pm 0.54}$	$97.35_{\pm 0.62}$	$97.36_{\pm 0.63}$	$97.36_{\pm 0.69}$
Amazon	$93.37_{\pm 1.18}$	$93.22_{\pm 1.58}$	$93.89_{\pm 1.27}$	$91.39_{\pm 2.03}$	$92.37_{\pm 1.54}$	$92.04_{\pm 1.58}$
YelpChi	$86.66_{\pm 0.60}$	$86.15_{\pm 0.60}$	$89.66_{\pm 0.57}$	$75.85_{\pm 1.06}$	$79.84_{\pm 1.26}$	$78.74_{\pm 1.29}$
T-Finance	$89.60_{\pm 0.49}$	$90.07_{\pm 0.42}$	$90.70_{\pm 0.59}$	$89.10_{\pm 0.80}$	$89.38_{\pm 0.58}$	$88.96_{\pm 0.63}$
Elliptic	$78.38_{\pm 0.00}$	$78.38_{\pm 0.00}$	$78.66_{\pm 0.00}$	$77.47_{\pm 0.34}$	$77.02_{\pm 0.34}$	$77.07_{\pm 0.28}$
Tolokers	$53.47_{\pm 0.96}$	$55.22_{\pm 0.66}$	$56.58_{\pm 1.09}$	$55.66_{\pm 0.92}$	$57.55_{\pm 1.02}$	$57.67_{\pm 1.14}$
Questions	$20.86_{\pm 2.01}$	$20.83_{\pm 1.74}$	$21.39_{\pm 1.79}$	$17.74_{\pm 1.77}$	$16.48_{\pm 1.81}$	$16.25_{\pm 1.93}$
DGraph-Fin	$3.83_{\pm 0.00}$	$4.20_{\pm 0.00}$	$4.19_{\pm 0.00}$	$2.51_{\pm 0.03}$	$2.96_{\pm 0.02}$	$2.78_{\pm 0.05}$
T-Social	$92.41_{\pm 0.10}$	$98.06_{\pm 0.03}$	$99.53_{\pm 0.03}$	$94.00_{\pm 0.06}$	$99.27_{\pm 0.03}$	$99.27_{\pm 0.02}$
Limited labels						
Reddit	$4.27_{\pm 0.40}$	$4.52_{\pm 0.53}$	$4.62_{\pm 0.72}$	$4.71_{\pm 0.64}$	$4.78_{\pm 0.64}$	$4.93_{\pm 0.81}$
Weibo	$87.63_{\pm 6.12}$	$88.56_{\pm 4.30}$	$88.02_{\pm 4.26}$	$93.63_{\pm 1.66}$	$92.07_{\pm 2.21}$	$91.38_{\pm 2.19}$
Amazon	$83.91_{\pm 1.75}$	$83.33_{\pm 3.26}$	$83.78_{\pm 3.42}$	$74.89_{\pm 3.57}$	$79.69_{\pm 6.03}$	$79.05_{\pm 4.89}$
YelpChi	$23.07_{\pm 3.90}$	$23.50_{\pm 3.75}$	$30.82_{\pm 6.20}$	$20.86_{\pm 2.11}$	$26.69_{\pm 4.22}$	$24.99_{\pm 4.03}$
T-Finance	$74.38_{\pm 7.89}$	$76.35_{\pm 2.89}$	$76.61_{\pm 4.53}$	$80.54_{\pm 2.91}$	$78.96_{\pm 3.08}$	$76.91_{\pm 3.94}$
Elliptic	$76.53_{\pm 4.41}$	$75.96_{\pm 5.83}$	$73.20_{\pm 5.32}$	$81.24_{\pm 3.17}$	$79.46_{\pm 3.71}$	$77.96_{\pm 5.16}$
Tolokers	$37.68_{\pm 4.47}$	$39.55_{\pm 3.93}$	$40.24_{\pm 3.57}$	$38.38_{\pm 2.37}$	$40.25_{\pm 2.85}$	$40.94_{\pm 3.48}$
Questions	$6.51_{\pm 1.98}$	$6.55_{\pm 1.79}$	$6.31_{\pm 1.39}$	$8.48_{\pm 2.35}$	$8.38_{\pm 1.99}$	$8.20_{\pm 2.00}$
DGraph-Fin	$2.16_{\pm 0.24}$	$2.20_{\pm 0.27}$	$2.19_{\pm 0.19}$	$2.24_{\pm 0.18}$	$2.34_{\pm 0.19}$	$2.26_{\pm 0.21}$
T-Social	$41.12_{\pm 5.49}$	$72.64_{\pm 7.94}$	$73.21_{\pm 9.33}$	$49.37_{\pm 4.21}$	$62.81_{\pm 6.10}$	$77.86_{\pm 4.31}$

Table 17: MSS-T reference controls: paired AUPRC changes over ten runs; positive values favor the control.

	Full supervision			Limited labels		
Reference	Weibo	Amazon	DGraph-Fin	Weibo	Amazon	DGraph-Fin
Materials only	$-0.07_{\pm 0.28}$	$-0.06_{\pm 0.41}$	$+0.16_{\pm 0.11}$	$+0.40_{\pm 1.73}$	$-0.91_{\pm 1.55}$	$+0.04_{\pm 0.20}$
Global reference	$+0.01_{\pm 0.28}$	$-0.07_{\pm 0.45}$	$+0.16_{\pm 0.15}$	$-0.31_{\pm 1.61}$	$-0.75_{\pm 1.06}$	$+0.07_{\pm 0.22}$
Neighbor reference	$-0.08_{\pm 0.19}$	$-0.17_{\pm 0.38}$	$+0.09_{\pm 0.11}$	$-1.51_{\pm 4.27}$	$-1.21_{\pm 2.02}$	$+0.01_{\pm 0.20}$

therefore holds 18 cells per graph and 1440 runs in total (Figure 9, Table 18). Across the eight graphs the shared default $(6, 0.3, 100)$ ranks between $6/18$ and $18/18$. The best fixed cell beats it by 2.93 AUPRC on Tolokers, 2.20 on Questions, 2.14 on YelpChi, 1.72 on Reddit, and 1.25 on Elliptic, while Weibo and Amazon stay within 0.26 points; selecting one cell per run by validation AUPRC and only then reading the test split recovers 2.68 and 2.13 points on Tolokers and Questions, so the headroom is reachable without test-set information. The largest losses occur at slow-learning cells: YelpChi $(4, 0.05, 100)$ loses 13.55 points and Elliptic $(8, 0.05, 100)$ loses 9.01, whereas no cell on the remaining six graphs falls more than 0.61 points below the default. Re-running the shared default inside the scan reproduces Table 1 within 0.11 AUPRC on every graph, so both sweeps describe the same operating point. These are descriptive test-set responses over a fixed grid; they do not select a replacement configuration for the main results.

Neural material removals. Figure 10 reports the seven-graph study. Each removed-space run is compared with its matched c-base run, not an unrelated main-table execution. This intervention removes a private neural branch, whereas the flat-tree study removes declared columns. Their different intervention units preclude treating the two figures as identical ablations of a shared numerical input.

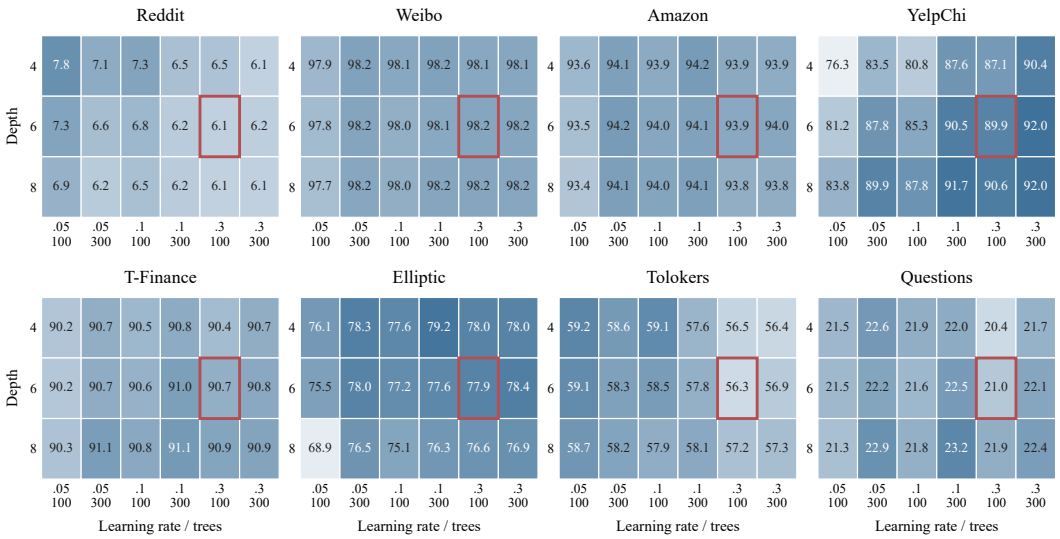

Figure 9: Hyperparameter landscape of the tree read-out: fully supervised AUPRC ($\times 100$, ten seeds) over depth (rows) $\times$ learning rate η and estimator count (columns). Each panel is shaded to its own range (darker is higher); numbers are cell means. The red box marks the shared default: depth 6, rate 0.3, and 100 trees.

Table 18: Grid response on the eight non-industrial graphs (full supervision, ten seeds, AUPRC $\times 100$). *Rank* places the shared default among the 18 swept cells (1 is best). Δ_{fix} is the best fixed cell minus the shared default; Δ_{val} selects one cell per run by validation AUPRC before the test split is read; *worst* is the lowest cell mean minus the shared default. Bold marks $\Delta_{\text{fix}} > 1$ AUPRC.

Dataset	Default	Rank	Best cell (d, η, n)	Δ_{fix}	Δ_{val}	Worst
Reddit	6.10	15/18	4, .05, 100	**+1.72**	+1.45	-0.04
Weibo	98.16	6/18	4, .1, 300	+0.06	+0.03	-0.44
Amazon	93.94	10/18	6, .05, 300	+0.26	-0.00	-0.58
YelpChi	89.89	7/18	6, .3, 300	**+2.14**	**+2.11**	-13.55
T-Finance	90.69	12/18	8, .1, 300	+0.44	+0.37	-0.52
Elliptic	77.95	7/18	4, .1, 300	**+1.25**	**+1.25**	-9.01
Tolokers	56.31	18/18	4, .05, 100	**+2.93**	**+2.68**	+0.00
Questions	20.98	17/18	8, .1, 300	**+2.20**	**+2.13**	-0.61

Nuisance-space control. A row-permuted individual-space view supplies a sixth expert with mismatched node evidence; an exact duplicate supplies an additional view without the permutation. Figure 11 compares both with the refreshed five-space run; the plotted paired differences contain the complete ten-dataset values. On Elliptic, the permuted view reduces AUPRC by 6.71, while the duplicate improves it by 5.04; on Questions the changes are -3.53 and -0.47. The duplicate itself can change predictions, so adding a slot cannot be assumed harmless. The difference between the two controls quantifies the replacement under this six-space architecture, rather than an independent causal effect of noise. Per-run AUPRC is stored to 0.01 points; differences near this precision should not be overinterpreted.

Label-budget scan. Both modes use the same frozen, nested stratified subsets of 100, 200, 400, and 800 nodes from each fully supervised training pool, with fixed validation and test masks and ten GADBench model seeds (Figure 12). MSS-T has higher mean AUPRC on YelpChi at every budget, whereas MSS-N leads on Questions. On T-Finance, MSS-N leads at 100 and 200 labels, and MSS-T at 400 and 800. Thus mode preference depends on both the dataset and label budget. Both modes' means rise with the budget; individual runs need not be monotone. The main 20+80 track has a different validation/test allocation and is reported separately.

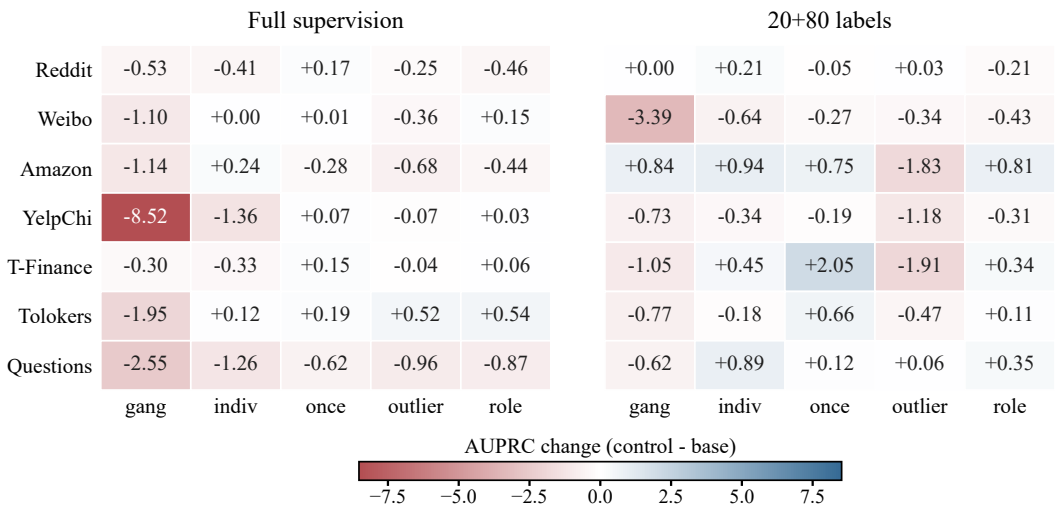

Figure 10: MSS-N space removal minus matched base, mean AUPRC change, ten runs. Both panels use the same signed color scale; blue/red indicate positive/negative changes.

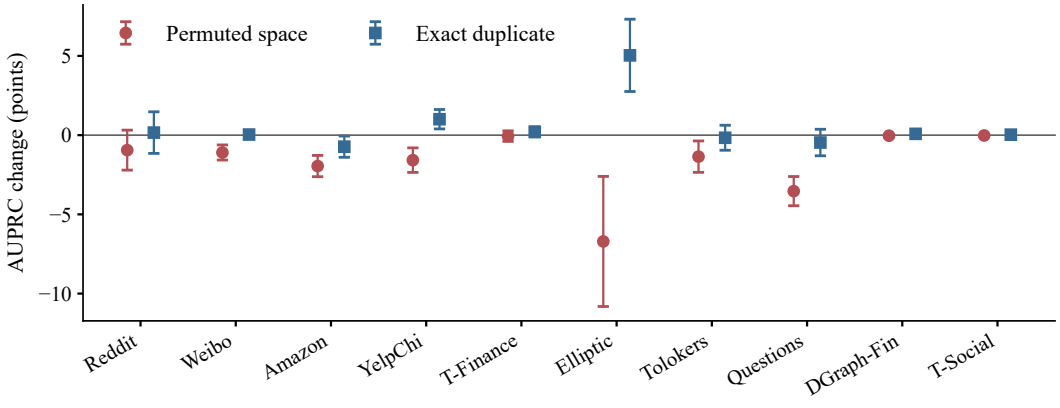

Figure 11: MSS-N nuisance and duplicate controls relative to the five-space base, full supervision. Means and 95% normal intervals over ten paired runs; control AUPRC is stored to 0.01 points per run.

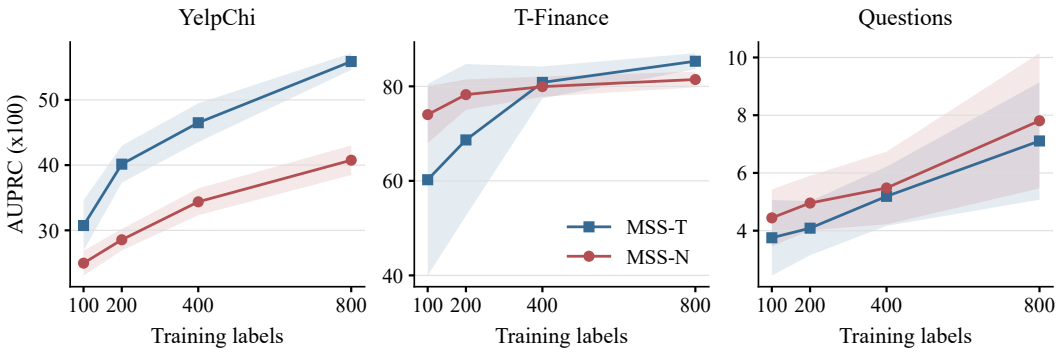

Figure 12: MSS-T and MSS-N label-budget scan: mean AUPRC and one-standard-deviation bands over ten GADBench model seeds. Both modes share training subsets and fixed validation/test masks.

K Efficiency: Runtime and Memory

Construction and retraining. Table 19(a) separates MSS-T's one-time construction from its per-seed supervised stage. Loading/materials includes graph loading and label-free material/backbone construction; references is the final evidence assembly. Both are shared single measurements from the limited-label main job, not ten independent construction trials. The supervised stage includes fitting, all-node scoring, and metric computation, summarized over ten runs. Seconds of tree fitting therefore do not imply seconds for the full pipeline; materials can be reused across fits on the same graph.

Merged-space cost. Table 19(b) reports merged MSS-N training time, parameter count, and peak allocated CUDA memory under both budgets. Training excludes model/graph loading and diagnostic exports; allocated memory differs from device-resident memory. The predictions in Table 10 also change expert count and objective organization. Timers with different stage boundaries cannot establish a paired speedup. Both main modes complete ten graphs under the single-card budget; Proposition 2 gives the structural memory argument.

Table 19: MSS resource costs: construction is measured once per job; fitting, training, and peak memory summarize ten runs.

(a) MSS-T: construction and supervised stage (seconds)

Dataset	Load/materials	References	Full fit/eval	Limited fit/eval
Reddit	0.47	0.05	$1.70_{\pm 0.14}$	$0.28_{\pm 0.09}$
Weibo	3.34	0.20	$3.50_{\pm 0.16}$	$0.51_{\pm 0.08}$
Amazon	18.10	0.02	$0.87_{\pm 0.05}$	$0.12_{\pm 0.06}$
YelpChi	16.44	0.10	$2.31_{\pm 0.05}$	$0.21_{\pm 0.06}$
T-Finance	40.07	0.03	$0.84_{\pm 0.05}$	$0.11_{\pm 0.07}$
Elliptic	4.55	2.10	$6.45_{\pm 0.45}$	$0.69_{\pm 0.08}$
Tolokers	2.19	0.01	$0.80_{\pm 0.00}$	$0.13_{\pm 0.07}$
Questions	1.78	0.86	$16.11_{\pm 0.24}$	$0.81_{\pm 0.08}$
DGraph-Fin	40.07	4.78	$7.92_{\pm 0.07}$	$1.34_{\pm 0.08}$
T-Social	393.87	4.62	$13.57_{\pm 0.07}$	$2.36_{\pm 0.15}$

(b) Merged-space MSS-N

		Training (seconds)		Peak allocated GiB	
Dataset	Params (M)	Full	Limited	Full	Limited
Reddit	0.37	$4.9_{\pm 0.0}$	$3.7_{\pm 0.0}$	$0.34_{\pm 0.00}$	$0.15_{\pm 0.00}$
Weibo	0.67	$18.6_{\pm 0.1}$	$15.7_{\pm 0.0}$	$1.29_{\pm 0.00}$	$1.29_{\pm 0.00}$
Amazon	0.34	$13.7_{\pm 0.0}$	$10.6_{\pm 0.1}$	$1.44_{\pm 0.00}$	$1.44_{\pm 0.00}$
YelpChi	0.35	$28.9_{\pm 0.1}$	$18.9_{\pm 0.2}$	$0.59_{\pm 0.00}$	$0.48_{\pm 0.00}$
T-Finance	0.33	$43.7_{\pm 0.8}$	$32.5_{\pm 0.3}$	$1.70_{\pm 0.03}$	$1.85_{\pm 0.10}$
Elliptic	0.47	$41.4_{\pm 0.1}$	$35.5_{\pm 0.1}$	$2.69_{\pm 0.00}$	$2.69_{\pm 0.00}$
Tolokers	0.33	$4.9_{\pm 0.0}$	$3.7_{\pm 0.0}$	$0.32_{\pm 0.00}$	$0.15_{\pm 0.00}$
Questions	0.58	$28.6_{\pm 0.2}$	$19.6_{\pm 0.1}$	$0.72_{\pm 0.00}$	$0.48_{\pm 0.00}$
DGraph-Fin	0.33	$536.4_{\pm 0.5}$	$253.4_{\pm 10.5}$	$3.52_{\pm 0.00}$	$3.50_{\pm 0.00}$
T-Social	0.33	$2220.6_{\pm 164.4}$	$636.1_{\pm 1.2}$	$5.73_{\pm 0.00}$	$5.68_{\pm 0.00}$

L Coordinate Diagnostics and Exported State

We record DGraph-Fin with temporal and directed spaces, and high-AUPRC Weibo, under both budgets at seed 3407: four diagnostic runs, each with 125 pre-update epoch states, a restored validation-selected state, and a PDF atlas. These cases illustrate the trained representation, not an additional benchmark average.

Before each diagnostic training run, a fixed earlier snapshot assigns test nodes to TP, FN, FP, TN, and near-threshold groups at probability 0.5. A fixed sampling RNG selects at most 32 nodes per group. This is post-hoc, label-assisted diagnostic selection: these labels do not enter training or

checkpoint choice. Figure 13 displays the smallest node ID in the predetermined TP and FN groups for each graph, without selecting for a visually favorable trajectory. Coordinates use the epoch-zero axis normalization throughout; score and reference changes are retained alongside motion. The same node follows different paths in different material spaces, and the fused probability need not change monotonically. This is the diagnostic information lost by a single merged coordinate system. Neither coordinate motion nor greater bin purity proves a real-world mechanism: interpretation requires the material view, occupancy, score, and contemporaneous reference. The archive includes all selected nodes, both budgets, coordinate arrays, occupancies, reference states, fusion weights, and the complete PDF atlases.

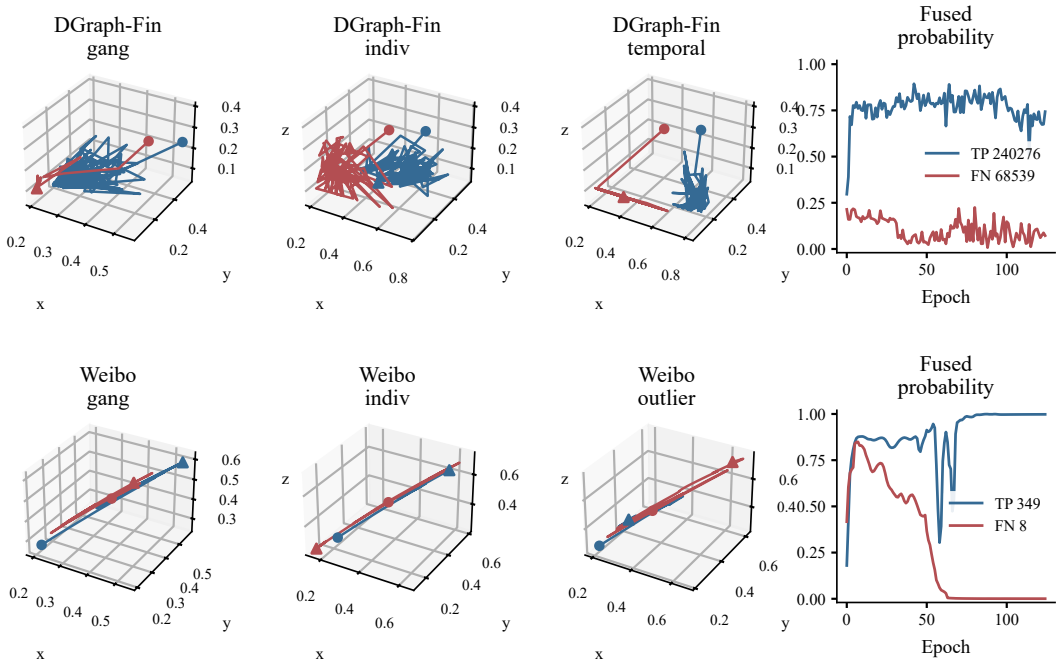

Figure 13: Fully supervised diagnostic trajectories in three selected spaces and fused probabilities. Circles/triangles mark the first/last recorded pre-update state; axes remain fixed to epoch zero.