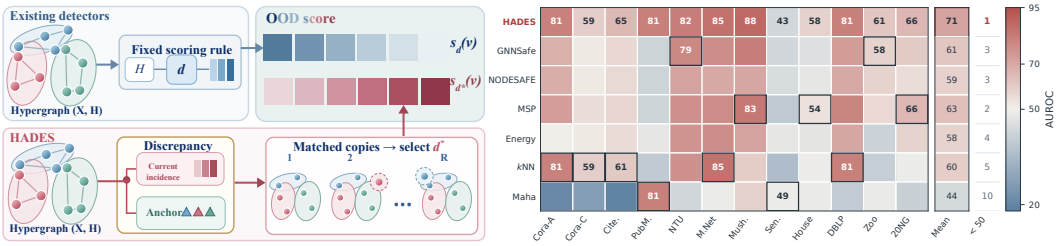
HADES: SELF-VALIDATED ENERGY SELECTION FOR STRUCTURAL OUT-OF-DISTRIBUTION DETECTION ON HYPERGRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hypergraphs represent higher order interactions among groups of entities, and hypergraph neural networks (HNNs) exploit these relations for prediction. At deployment, hyperedge formation may change, placing nodes in relational contexts unseen during training. Structural out of distribution (OOD) detection flags when confident predictions rely on relational patterns no longer supported by training, making it essential for reliable deployment. Yet this task remains underexplored, and no fixed post hoc score is reliable across datasets and backbones. We introduce HADES (**H**ypergraph **A**daptive **D**etection via **E**nergy **S**election), which pairs a new structural score with adaptive detector selection. The score measures disagreement between log partition confidence propagated over the current incidence and a soft incidence anchored by prototypes fitted to clean training embeddings. To handle detector heterogeneity, HADES turns score choice into a predeployment self validation problem. Matched perturbations of the clean hypergraph define a surrogate shift. Using this surrogate, HADES selects and freezes the single scoring rule best aligned with anticipated deployment changes. Across 12 hypergraphs and four backbones, HADES achieves the best average detector rank on every backbone. Its overall average rank is 2.16, compared with 5.82 for runner up KLM. These results show that reliable structural OOD detection depends on both a clean relational reference and score selection aligned with the anticipated shift.



(a) Method comparison **(b) Performance across datasets**

Figure 1: **HADES couples a relational reference with detector selection.** *(a)* HADES contrasts the current incidence with a prototype anchor and selects one detector using matched structural copies. *(b)* Structural OOD AUROC on 12 datasets with HGNN. Values mark HADES and the outlined best fixed detector among those displayed for each dataset. The right columns report mean AUROC and counts below 50.

1 INTRODUCTION

Hypergraphs represent higher-order relations through memberships shared by groups of nodes, and hypergraph neural networks (HNNs) use these memberships to propagate information for prediction (Feng et al., 2019; Huang & Yang, 2021; Chien et al., 2022). At deployment, however, the incidence structure may change even when the node set and features remain fixed. New collaborations may

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.

form, existing groups may dissolve, and nodes may enter relational contexts that were absent during training. An HNN can still assign high confidence to predictions formed through these unfamiliar relations. Structural out-of-distribution (OOD) detection is therefore a safeguard on the relational basis of a prediction: it asks when that supporting context has left the training regime.

Recent OOD detectors draw on classifier outputs or feature geometry (Hendrycks et al., 2022; Liu et al., 2023; Sun et al., 2022; Wang et al., 2022; Park et al., 2023). Graph-specific methods such as GNNSafe, NODESAFE, ML-GOOD, and the recent UniGOD additionally exploit relational structure or shift-dependent geometry (Wu et al., 2023; Yang et al., 2024; Cai et al., 2025; Song et al., 2026). We adapt GNNSafe and NODESAFE to hypergraph propagation. The most directly relevant prior work is HyperGOOD, which formalizes OOD detection for hypergraphs through multiscale framelet representations and structure-aware energy propagation (Cai et al., 2026; Li et al., 2025). Its evaluation, however, perturbs features and labels while keeping the incidence fixed, so it does not address how to construct a separate relational reference when the incidence itself changes. Figure 1(a) distills the resulting gap: a fixed rule scores the current hypergraph alone, whereas our two-part design introduces a relational anchor and selects among complementary rules before deployment using matched structural copies.

The relational reference addresses incidence comparison, while detector heterogeneity poses a second challenge. Structural shifts may appear as low confidence, atypical features, or inconsistent propagation, and the most informative signal therefore varies across datasets and backbones. In our benchmark, every one of the 13 existing fixed detectors falls below random AUROC on at least two of 12 hypergraphs with HGNN. Figure 1(b) shows six representative detectors, and Appendix C.4 reports the full comparison. The lack of a clean relational reference and a consistently reliable detector therefore motivates two capabilities: a discrepancy between current-incidence propagation and an anchor defined by clean prototypes, and self-validation over surrogate perturbations chosen to represent plausible deployment changes.

We introduce Hypergraph Adaptive Detection via Energy Selection (HADES) to realize this two-part design. Its first component constructs a prototype-anchored energy discrepancy. Prototypes are fitted once to clean training embeddings and then kept fixed. Current node embeddings are softly assigned to these prototypes, producing a reference incidence that is not built from the current hyperedges. HADES propagates the same log-partition confidence through the current and prototype-anchored incidences, then measures their standardized disagreement. A structural change enters the current-incidence view directly. It can affect the prototype-anchored view only through the frozen backbone outputs and the prototype bottleneck. Their discrepancy therefore supplies evidence that the current relational context is inconsistent with the clean representation geometry.

Because structural shifts favor different signals across settings, HADES places the discrepancy alongside seven established detectors as alternative detection hypotheses. Before deployment, matched structural perturbations of the clean hypergraph create a surrogate task on held out validation nodes, with the provenance of clean and perturbed copies supplying the ID/OOD labels. HADES selects the single detector with the best surrogate AUROC and freezes it before target evaluation (Figure 2). It neither averages detector outputs nor adapts on target scores, and it requires no target OOD labels. The accompanying regret analysis gives a sufficient condition under which self validation approaches oracle selection, namely that surrogate detector performance uniformly approximates performance under the target shift.

Across 12 hypergraphs and four HNN backbones, HADES achieves the best average rank on every backbone and improves overall mean AUROC by 7.1 points over runner-up KLM, from 60.8 to 67.9. Self-validation is evaluated independently in all three runs for every dataset and backbone setting. On the ten datasets for which our HyperGOOD reproduction completes on the stated 16 GB GPU, HADES leads HyperGOOD by 8.1 mean AUROC points. On the other two, clique-induced operator products exceed our pre-specified resource screen (Appendix A.5). Against our hypergraph adaptation UniGOD-H, HADES is higher on 11 of 12 datasets under the same structural protocol (Appendix A.4). The ranking advantage also persists across three independently generated structural copies. These results cast structural OOD detection as a joint problem of relational reference construction and signal selection matched to plausible deployment shifts, rather than a search for one universally reliable score.

2 Related Work

Post-hoc OOD scores. Post-hoc OOD scores support reliable deployment by indicating when an input may lie outside the training distribution. They can be computed from a trained model without changing it and typically use either classifier outputs or learned features. MSP, ODIN, Energy, MaxLogit, KL-Matching, and GEN use confidence or logit statistics (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Hendrycks et al., 2022; Liu et al., 2023). Mahalanobis, Relative-Mahalanobis, kNN, ViM, and NNGuide use distances or residual geometry in the representation space (Lee et al., 2018; Ren et al., 2021; Sun et al., 2022; Wang et al., 2022; Park et al., 2023). When anomaly labels are unavailable, MetaOD transfers detector performance from historical tasks, while surrogate criteria based on injected anomalies support detector selection for time series (Zhao et al., 2021; Goswami et al., 2023). HADES instead uses matched incidence perturbations within the clean hypergraph to select a nodewise structural OOD score. These methods provide complementary confidence- and geometry-based signals, but they do not explicitly model shifts in hypergraph incidence.

Graph and hypergraph OOD detection. Reliable OOD detection is especially important for graph and hypergraph models because a structural change can affect multiple nodes through message passing. The same coupling makes detection difficult because both predictions and structure-aware scores depend on the relational structure being assessed. Model-driven reconstruction offers a complementary reliability perspective: convergent unrolling preserves an explicit iterative operator, while uncertainty-principle analysis relates graph-signal recoverability to incomplete observations and noise (Zhao et al., 2024; 2025). GNNSafe (Wu et al., 2023) propagates log-partition confidence over a graph, while NODESAFE (Yang et al., 2024) adds a bounded-energy training regularizer. ML-GOOD (Cai et al., 2025) addresses multi-label node OOD on graphs. Recent graph methods broaden the setting: UniGOD adapts node-representation geometry across diverse shifts, whereas BaCa calibrates graph-level scores at test time and PGOS synthesizes graph-level outliers (Song et al., 2026; Hou et al., 2026; Sun et al., 2026).

HyperGOOD (Cai et al., 2026) is the closest hypergraph counterpart. It combines framelet-based representations (Li et al., 2025) with structure-aware energy propagation, but evaluates feature and label perturbations under a fixed incidence. A separate line of hypergraph research learns or refines incidence for predictive accuracy through dynamic neighborhoods, explicit structure optimization, or attention-refined Laplacians (Jiang et al., 2019; Cai et al., 2022; Zhang et al., 2022). HADES instead treats incidence shift as the detection target and uses a prototype incidence only as a detector reference, leaving the predictive backbone unchanged.

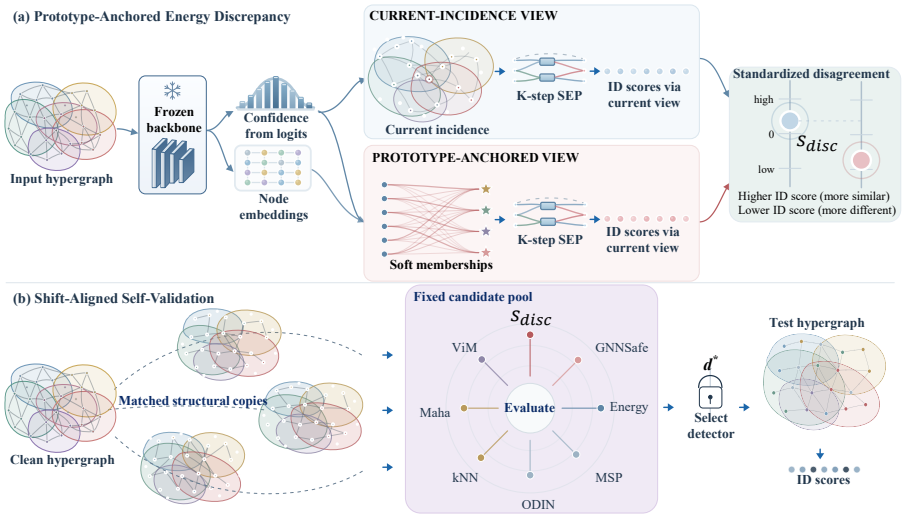


Figure 2: **HADES overview.** *(a)* The prototype-anchored discrepancy compares shared confidence under the current and clean-prototype incidences. *(b)* Matched structural copies rank candidates on held-out nodes before one is frozen for test scoring.

3 METHOD

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{H}, \mathbf{X})$ denote the clean hypergraph used to train a backbone HNN f for C-class ID node classification. Here $\mathcal{V}$ is the set of N nodes, and $\mathcal{E}$ is the set of M clean hyperedges. The incidence $\mathbf{H} \in \{0,1\}^{N\times M}$ encodes their memberships, and $\mathbf{X} \in \mathbb{R}^{N\times F}$ contains the F-dimensional node features. Let $\mathcal{V}_{\text{tr}}$ and $\mathcal{V}_{\text{val}}$ denote the fixed training and validation node indices. At deployment, the detector receives a hypergraph $\widetilde{\mathcal{G}} = (\mathcal{V}, \widetilde{\mathcal{E}}, \widetilde{\mathbf{H}}, \mathbf{X})$ that retains $\mathcal{V}$ and $\mathbf{X}$ but may have a different hyperedge set $\widetilde{\mathcal{E}}$. Writing $\widetilde{M} = |\widetilde{\mathcal{E}}|$, its incidence matrix $\widetilde{\mathbf{H}} \in \{0,1\}^{N\times\widetilde{M}}$ encodes $\widetilde{\mathcal{E}}$ and may differ from $\mathbf{H}$. A post-hoc OOD detector assigns every node $v \in \mathcal{V}$ a score $s(v \mid \mathbf{X}, \widetilde{\mathbf{H}}) \in \mathbb{R}$. Larger values indicate that the node's deployment context remains consistent with the clean regime. For a structure-aware score, however, $\widetilde{\mathbf{H}}$ defines both the deployment context being assessed and the propagation operator used to score it. It therefore cannot also provide the required clean relational reference.

HADES first constructs a *Prototype Anchored Energy Discrepancy* (Section 3.1) by comparing confidence propagated over a supplied incidence with confidence propagated over a reference fitted to clean training representations. This construction supplies a structural signal for detecting when the current relational context departs from the clean regime. To remain reliable when a shift is expressed more strongly by complementary evidence, HADES then performs *Shift Aligned Self Validation* (Section 3.2) over a fixed signal pool before $\widetilde{\mathbf{H}}$ is scored.

3.1 PROTOTYPE-ANCHORED ENERGY DISCREPANCY

One signal, two relational views. A confidence score propagated only through $\widetilde{\mathbf{H}}$ is conditioned on the same incidence being assessed and therefore lacks a clean relational comparator. To reveal whether the target relationships still support the node-wise evidence, HADES compares one confidence vector under current-incidence and prototype-anchored views. Sharing the confidence signal and propagation rule prevents their disagreement from being explained by different scoring functions. The frozen backbone f evaluated on $(\mathbf{X}, \widetilde{\mathbf{H}})$ returns a logit matrix $\widetilde{\mathbf{Z}} \in \mathbb{R}^{N\times C}$ and a penultimate embedding matrix $\widetilde{\mathbf{h}} \in \mathbb{R}^{N\times d}$, where d is the embedding dimension. For node v, $\widetilde{Z}_{vc}$ denotes its class-c logit and $\widetilde{\mathbf{h}}_v$ denotes its row in the embedding matrix. For temperature $T > 0$, the scalar log-partition confidence of node v is

$$\widetilde{q}_v = T \log \sum_{c=1}^{C} \exp\left(\widetilde{Z}_{vc}/T\right), \tag{1}$$

which is the negative free energy and is larger for ID under the usual energy convention (Liu et al., 2020; Grathwohl et al., 2020). Stacking these values gives $\widetilde{\mathbf{q}} = (\widetilde{q}_v)_{v\in\mathcal{V}} \in \mathbb{R}^N$.

A clean relational anchor. The second view must preserve group geometry learned from the clean regime, assign every target node to that reference, and avoid reconstructing the target hyperedges. We meet these requirements with fixed clean prototypes and dynamic node assignments. Evaluating f on $(\mathbf{X}, \mathbf{H})$ gives the clean embedding matrix $\mathbf{h} \in \mathbb{R}^{N\times d}$, whose row $\mathbf{h}_v$ represents node v. We write $\mathbf{u}_v = \mathbf{h}_v/\|\mathbf{h}_v\|_2$ and $\widetilde{\mathbf{u}}_v = \widetilde{\mathbf{h}}_v/\|\widetilde{\mathbf{h}}_v\|_2$ for the normalized clean and target embeddings. Spherical k-means fits K' centers $\{\boldsymbol{\mu}_j \in \mathbb{R}^d\}_{j=1}^{K'}$ to $\{\mathbf{u}_v : v \in \mathcal{V}_{\text{tr}}\}$, and these centers are then cached. For assignment temperature $\tau > 0$, the soft prototype incidence $\widehat{\mathbf{H}} \in [0,1]^{N\times K'}$ has entry

$$\widehat{\mathbf{H}}_{vj} = \frac{\exp(\langle \widetilde{\mathbf{u}}_v, \boldsymbol{\mu}_j\rangle/\tau)}{\sum_{r=1}^{K'} \exp(\langle \widetilde{\mathbf{u}}_v, \boldsymbol{\mu}_r\rangle/\tau)}. \tag{2}$$

Thus, $\widehat{\mathbf{H}}_{vj}$ is the soft membership of target node v in clean prototype j. The fixed centers provide a clean coordinate system, while the assignments locate target nodes within it. Consequently, $\widehat{\mathbf{H}}$ can respond to the target representations without being constructed from the memberships in $\widetilde{\mathbf{H}}$.

Matched propagation and discrepancy. With both incidences defined, HADES propagates the same confidence vector through each one using the same rule. For an incidence matrix $\mathbf{H}$, let $\mathbf{D}_v(\mathbf{H})$

and $\mathbf{D}_e(\mathbf{H})$ denote its diagonal node- and hyperedge-degree matrices. The normalized hypergraph operator is

$$P(\mathbf{H}) = \mathbf{D}_v(\mathbf{H})^{-1/2}\mathbf{H}\mathbf{D}_e(\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{D}_v(\mathbf{H})^{-1/2}. \tag{3}$$

We use the diagonal pseudoinverse convention $0^{-1/2} := 0$, so isolated nodes contribute zero through $P(\mathbf{H})$. For an integer $K \geq 1$, mixing coefficient $\alpha \in [0, 1]$, and input vector $\mathbf{q} \in \mathbb{R}^N$, the K-step structure-aware energy propagation (SEP) is

$$\mathrm{SEP}_{\mathbf{H}}(\mathbf{q}) = \mathbf{q}^{(K)}, \qquad \mathbf{q}^{(t+1)} = \alpha\mathbf{q}^{(t)} + (1-\alpha)P(\mathbf{H})\mathbf{q}^{(t)}, \quad \mathbf{q}^{(0)} = \mathbf{q}, \tag{4}$$

where $t = 0, \ldots, K-1$. This is the propagation used by GNNSafe and HyperGOOD (Wu et al., 2023; Cai et al., 2026). The current-incidence and prototype-anchored views are therefore $\mathrm{SEP}_{\widetilde{\mathbf{H}}}(\widetilde{\mathbf{q}})$ and $\mathrm{SEP}_{\widehat{\mathbf{H}}}(\widetilde{\mathbf{q}})$, respectively.

Let $z_{\mathrm{tr}}(\mathbf{a})$ standardize a vector $\mathbf{a}$ using its mean and standard deviation on $\mathcal{V}_{\mathrm{tr}}$. HADES defines the higher-is-ID score

$$s_{\mathrm{disc}}(v) = -\left|z_{\mathrm{tr}}\big(\mathrm{SEP}_{\widetilde{\mathbf{H}}}(\widetilde{\mathbf{q}})\big)_v - z_{\mathrm{tr}}\big(\mathrm{SEP}_{\widehat{\mathbf{H}}}(\widetilde{\mathbf{q}})\big)_v\right|. \tag{5}$$

Standardization removes view-specific scale, leaving their disagreement as the relevant quantity. A structural change enters the current-incidence view directly through $P(\widetilde{\mathbf{H}})$, but reaches the prototype-anchored view only through the frozen backbone outputs and the assignments of its embeddings to fixed clean centers. The views should therefore agree when the target context remains compatible with the clean regime and separate when the target incidence introduces a non-cancelling relational change.

Appendix B.1 formalizes this intuition. Let $D_{\mathbf{H}}(v)$ denote the unsigned form of the discrepancy score when it is evaluated on incidence $\mathbf{H}$. Here β is the clean approximation error, η_v is anchor drift, and $\rho_v\Gamma_v$ is the non-cancelling contamination effect formalized in Appendix B.1.

Proposition 1 (Discrepancy separation)**.** *Under Assumption 1 in Appendix B.1, a clean node satisfies $D_{\mathbf{H}}(v) \leq 2\beta$, whereas a structurally shifted node evaluated under incidence $\widetilde{\mathbf{H}}$ satisfies*

$$D_{\widetilde{\mathbf{H}}}(v) \geq \rho_v\Gamma_v - (2\beta + \eta_v). \tag{6}$$

Consequently, the clean and shifted discrepancy ranges are separated whenever $\rho_v\Gamma_v > 4\beta + \eta_v$.

This separation condition motivates retaining complementary signals in the detector pool. Appendix B.1 provides the assumptions and proof.

3.2 Shift-Aligned Self-Validation

The prototype anchored discrepancy makes incidence change directly measurable against a clean relational reference. Yet a structural shift may surface more strongly in classifier confidence or feature geometry, as Figure 1(b) shows across hypergraphs and backbones. HADES therefore treats its new discrepancy and established scores as complementary evidence channels and uses clean data perturbations to estimate which channel is reliable for the anticipated shift. This self validation produces a frozen, shift aligned detection rule before the target hypergraph is accessed.

We instantiate a fixed pool spanning structure-aware, classifier-output, and feature-geometry signals:

$$\mathcal{D} = \{s_{\mathrm{disc}}, \mathrm{GNNSafe}, \mathrm{Energy}, \mathrm{MSP}, \mathrm{ODIN}, k\mathrm{NN}, \mathrm{Mahalanobis}, \mathrm{ViM}\}.$$

Surrogate ranking without target access. Without target OOD labels, the target AUROC needed to rank the candidates is unavailable. HADES instead constructs a surrogate selection task entirely from the clean hypergraph using a supplied perturbation family that represents plausible deployment changes. The generator produces R surrogate incidences $\mathbf{H}^{(1)}_{\mathrm{sur}}, \ldots, \mathbf{H}^{(R)}_{\mathrm{sur}}$, and each candidate d returns a vector of higher-is-ID scores on held-out validation indices $\mathcal{V}_{\mathrm{val}}$:

$$\widehat{R}(d) = \frac{1}{R}\sum_{r=1}^{R} \mathrm{AUROC}_{\mathcal{V}_{\mathrm{val}}}\Big(d(\mathbf{X}, \mathbf{H}), d(\mathbf{X}, \mathbf{H}^{(r)}_{\mathrm{sur}})\Big), \qquad d^\star = \arg\max_{d\in\mathcal{D}} \widehat{R}(d). \tag{7}$$

Taking the clean copy as the positive class, $\widehat{R}(d)$ measures how well each signal separates clean from plausibly shifted relational contexts on held-out nodes. Self-validation therefore converts prior

knowledge of plausible incidence changes into a target-independent reliability estimate for each signal rather than selecting a detector on the target hypergraph. For the discrepancy candidate, Eq. (5) is instantiated separately on $\mathbf{H}$ and each surrogate using the logits, embeddings, and assignments produced by that incidence, while the clean prototype centers remain fixed. The maximizer $d^\star$ is then cached as the shift-aligned detection rule, and the target hypergraph is scored only once as $d^\star(\mathbf{X}, \widetilde{\mathbf{H}})$.

Appendix B.1 shows that uniform surrogate error ε yields selection regret at most 2ε. Algorithm 1 and all fixed settings are in Appendix A.2.

4 Experiments

4.1 Experimental Setup

Datasets and backbones. We evaluate 12 hypergraph benchmarks spanning relational, visual, categorical, text, and political-committee domains. The relational group contains Cora-CA, Cora-CC, Citeseer, PubMed, and DBLP. The remaining datasets are NTU2012, ModelNet40, Mushroom, Senate, House, Zoo, and 20News. We use the train, validation, and test protocol specified in Appendix A.1. The split is generated once with seed 42 and shared by every method and run. The training indices fit the backbone, the validation indices select its checkpoint and support target-free self-validation, and the test indices are used only for final evaluation. Four backbones cover spectral, residual, set based, and diffusion HNNs: HGNN (Feng et al., 2019), UniGCNII (Huang & Yang, 2021), AllSetTransformer (Chien et al., 2022), and ED-HNN (Wang et al., 2023). Dataset sources, split sizes, and full statistics are provided in Appendix A.1.

Structural OOD protocol. Each backbone is trained only on the clean hypergraph and then frozen before any structural OOD copy is scored. To construct a perturbed incidence $\widetilde{\mathbf{H}}$, each clean hyperedge is retained independently with probability $1 - p_{\mathrm{rm}}$. If M_{keep} hyperedges survive, we add $\lfloor p_{\mathrm{add}} M_{\mathrm{keep}} \rfloor$ new hyperedges, sampling each size from the surviving edge-size multiset and its incident nodes uniformly without replacement. The node identities, features, class labels, and split indices remain fixed, so each test node is evaluated once in its clean relational context and once under the perturbed incidence. Only the same fixed test indices enter the two final score sets. Training and validation nodes are excluded from final ID/OOD metrics. Main results use $p_{\mathrm{rm}} = 0.7$ and $p_{\mathrm{add}} = 0.5$, and every method within a dataset and backbone comparison receives the same copy in each run. This paired construction isolates incidence shift from changes in node content or population. Appendix B.2 evaluates independent perturbation copies, while Section 4.5 varies removal severity.

Compared methods. We compare HADES with 13 fixed detectors drawn from three signal families. Classifier-output methods comprise MSP, Energy, ODIN, MaxLogit, GEN, and KL-Matching (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Hendrycks et al., 2022; Liu et al., 2023). Feature-geometry methods comprise Mahalanobis, Relative-Mahalanobis, kNN, ViM, and NNGuide (Lee et al., 2018; Ren et al., 2021; Sun et al., 2022; Wang et al., 2022; Park et al., 2023). The relational baselines are GNNSafe and NODESAFE (Wu et al., 2023; Yang et al., 2024). All fixed detectors except NODESAFE reuse the same standard-trained backbone as HADES, while NODESAFE retrains each backbone with its required bounded-energy objective. The eight signal pool in Eq. (7) is fixed before target evaluation and spans relational confidence, classifier outputs, and feature geometry with training compatible representatives. We reproduce HyperGOOD under our structural protocol because it is the only peer-reviewed method explicitly designed for hypergraph OOD detection, although its original evaluation perturbs features and labels under fixed incidence (Cai et al., 2026). We also adapt UniGOD to hypergraphs as UniGOD-H because it targets node-level OOD under structure manipulation, making it the closest recent graph-side comparator (Song et al., 2026). Both replace the backbone and are therefore reported as separate model-level comparisons rather than detectors available to HADES. Appendix A.3 provides implementation details and summarizes how recent graph OOD methods differ in task definition.

Evaluation protocol. We report AUROC, AUPR, and FPR95 as percentages, treating clean-context test nodes as the positive ID class (Hendrycks & Gimpel, 2017). For every dataset, the ID and OOD score sets contain the same test nodes under clean and perturbed incidences, respectively. They are therefore equal in size, and random AUPR is 50%. AUROC is the primary ranking metric,

AUPR summarizes precision and recall behavior, and FPR95 is the false-positive rate when 95% of ID nodes are accepted. Backbone checkpoints are selected solely by clean ID validation accuracy. Unless noted, rankings include HADES and all 13 fixed detectors. Clean ID accuracy and complete per-metric results are reported in the appendix.

For every dataset, backbone, and run, HADES uses $R = 3$ target-independent surrogate copies, $K = 2$ SEP steps, $\alpha = 0.5$, $\tau = 0.5$, and $K' = \max(8, \min(64, 4C))$. The surrogate copies are generated independently of the target structural copy and are evaluated only on the fixed validation indices. Their clean-versus-perturbed provenance supplies the proxy ID/OOD labels, so neither target OOD labels nor target scores enter detector selection. All HADES settings are fixed globally rather than tuned per dataset or backbone. Appendix A.2 gives the remaining implementation details and computational resources. Unless stated otherwise, all stochastic experiments use global master seeds 42, 123, and 456. Model initialization and training use the corresponding master seed, while target perturbations and HADES surrogate copies use deterministic, non-overlapping random streams derived from it. Within each run, every compared method receives the same target copy, whereas HADES uses three distinct surrogate copies generated independently of that target. Any single-run visualization uses the run with master seed 42.

4.2 Detection Performance

We first test whether HADES improves structural OOD detection across heterogeneous hypergraphs and signal families. Table 1 evaluates HADES and six representative fixed detectors with the HGNN backbone, spanning relational, classifier-output, and feature-geometry scores, while the two model-level baselines retain their reproduced backbones. Complete per-dataset results for all 13 fixed detectors, four backbones, and three metrics are provided in Appendix C.4.

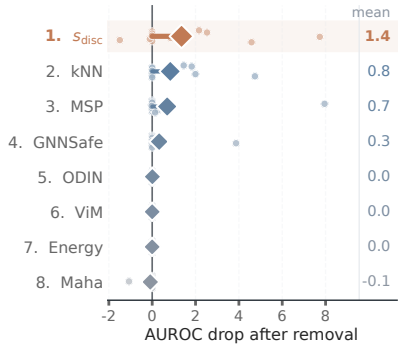
Table 1: **Structural OOD detection** ($\times 100$). Dataset rows report AUROC over three runs. Summary rows add AUPR and FPR95. Bold and underlining use unrounded means to mark the best and second best results among methods covering all 12 datasets. HyperGOOD† reports its ten completed datasets and is not included in these markers.

	Hypergraph / graph OOD					*Transferred generic*			
Dataset	**HADES**	**HyperGOOD**	**UniGOD-H**	**GNNSafe**	**NODESAFE**	**GEN**	**MSP**	**Maha**	**kNN**
Cora-CA	**80.8±3.2**	72.9±0.6	68.4±2.3	67.8±2.3	68.4±3.3	71.7±1.5	72.2±1.3	21.1±1.2	**80.8±3.2**
Cora-CC	**59.4±1.9**	54.9±0.9	53.9±0.3	52.0±1.0	51.9±0.9	54.1±0.7	54.9±0.7	26.9±1.1	59.2±1.7
Cite.	**65.3±0.8**	57.2±0.4	44.3±0.7	56.8±0.4	56.1±0.5	57.1±0.4	57.3±0.4	19.7±1.1	60.6±0.8
PubM.	81.1±1.4	42.2±0.5	57.9±0.5	42.1±0.5	42.1±0.5	42.2±0.5	42.2±0.5	**81.1±0.1**	39.4±0.6
NTU	**81.8±1.0**	77.3±0.8	55.3±2.7	78.7±0.5	71.4±1.4	78.7±0.7	76.4±0.8	42.9±1.3	76.1±1.4
M.Net	**85.2±0.7**	75.1±0.3	56.4±1.0	77.3±1.8	69.6±0.7	78.5±0.1	77.7±0.2	48.7±0.7	85.1±0.0
Mush.	**87.6±4.6**	N/A	49.1±3.0	81.5±13.1	70.4±5.2	83.4±2.0	83.4±2.0	58.1±2.3	59.9±7.3
Sen.	43.1±6.3	44.2±2.3	**50.6±0.6**	43.2±3.5	44.5±2.0	38.9±4.8	39.0±4.9	48.8±0.2	35.8±8.0
House	**58.2±3.1**	47.8±0.9	52.4±0.3	45.6±1.6	47.5±1.4	54.0±2.8	54.0±2.8	48.4±1.9	49.1±9.3
DBLP	**81.2±0.6**	77.5±0.2	69.8±0.6	73.5±0.6	69.4±0.5	80.2±0.2	79.6±0.2	40.1±1.0	**81.2±0.6**
Zoo	61.3±14.3	67.5±11.0	48.4±1.5	57.8±12.1	57.8±12.1	**68.5±20.8**	56.6±8.8	48.3±4.4	40.8±19.1
20NG	65.7±1.2	N/A	54.5±4.2	60.5±4.0	55.6±3.8	65.3±1.3	**65.8±1.1**	41.3±5.3	48.9±4.9
Mean AUROC↑	**70.9**	61.7†	55.1	61.4	58.7	64.4	63.3	43.8	59.7
Mean AUPR↑	**72.5**	62.5†	54.2	60.0	57.9	62.8	62.0	51.6	62.6
Mean FPR95↓	**77.9**	82.5†	89.7	80.9	85.9	78.4	80.4	95.5	85.4

Overall detection quality. Using the HGNN backbone, HADES reaches 70.9 mean AUROC, 6.5 points above the strongest displayed fixed detector GEN, and is best or tied best among the displayed methods in 8 of the 12 dataset rows in Table 1. It exceeds HyperGOOD on 8 of the 10 rows for which that reproduction completes and leads UniGOD-H on 11 of 12 datasets, with a 15.8-point advantage in mean AUROC. These dataset specific differences reflect where an incidence perturbation becomes visible after message passing. On PubMed, the signal remains in class-conditional embedding geometry: HADES and Mahalanobis reach 81.1, while GEN and GNNSafe remain near 42. Cora-CA shows that feature geometry itself can favor local over class-conditional structure: HADES and kNN reach 80.8, whereas Mahalanobis reaches 21.1. On Zoo, the ordering reverses again, with GEN at 68.5, HADES at 61.3, and kNN at 40.8, so classifier-output statistics are more informative than local embedding density. These crossovers explain why no fixed detector is uniformly reliable and why shift-aligned selection drives HADES's aggregate advantage.

Table 2: **Component and candidate pool ablations** using the HGNN backbone. Left: AUROC mean±standard deviation over three runs for the component variants on four representative datasets. Right: three run mean AUROC change after removing one candidate. Dots are datasets and diamonds are means over all 12 datasets. Positive values indicate that removal hurts HADES.

Dataset	disc only	select w/o s_{disc}	full
Cora-CA	61.3±14.9	**76.0±1.0**	**76.0±1.0**
Citeseer	**65.8±0.1**	58.1±0.1	**65.8±0.1**
NTU2012	**80.1±0.1**	77.6±0.1	**80.1±0.1**
Mushroom	65.6±16.4	91.6±6.1	**92.2±5.2**
Mean	68.2	75.8	**78.5**



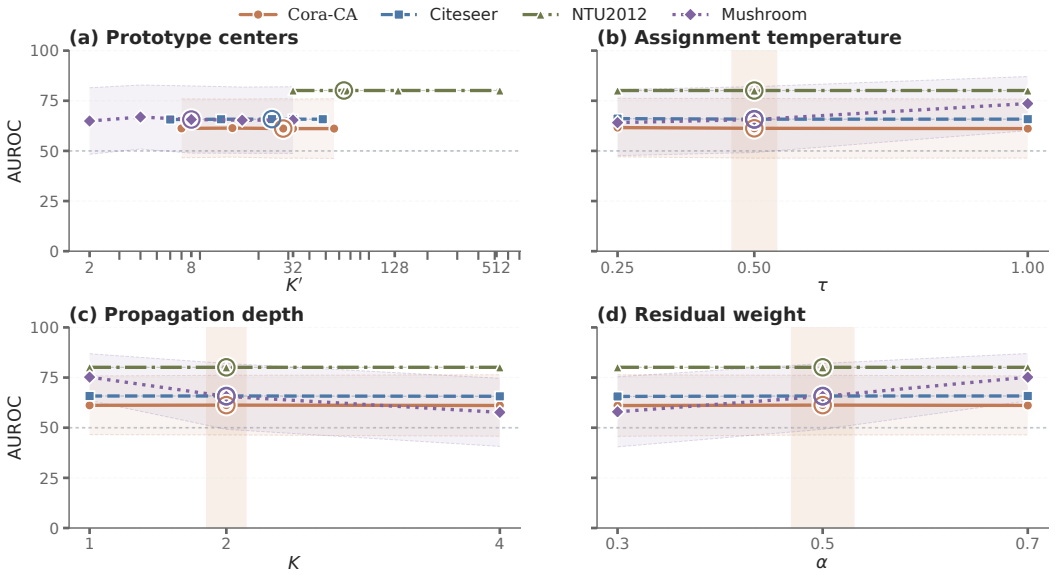


Figure 3: **Hyperparameter sensitivity of the discrepancy** using the HGNN backbone. Curves and bands show mean±1 standard deviation over three runs. White rings mark the fixed settings, and the horizontal dotted line marks random performance.

4.3 Selection and Component Ablations

We next test whether the aggregate gains arise from reliable candidate selection and complementary components. Table 2 reports a component comparison and the effect of removing each candidate, while Appendix B.4 compares the selector with an oracle selector that chooses the best candidate in the fixed pool using target labels. Across 48 dataset and backbone settings and three runs, the 144 selections span all eight candidates and remain close to oracle selection, indicating stable selection quality across seeds. Matching occurs when the full and reduced variants select the same signal, since HADES deploys one detector rather than combining scores. In Table 2 (left), the full method averages 78.5 AUROC, versus 75.8 without the discrepancy and 68.2 with it alone. The discrepancy is decisive on Citeseer and also supplies the strongest signal on NTU2012, whereas other candidates match or nearly match the full method on Cora-CA and Mushroom. In the right panel, removing s_{disc} causes the largest mean loss across 12 datasets (1.4 AUROC). The selector provides most of the gain over the discrepancy alone, while the discrepancy contributes the largest marginal improvement within the pool.

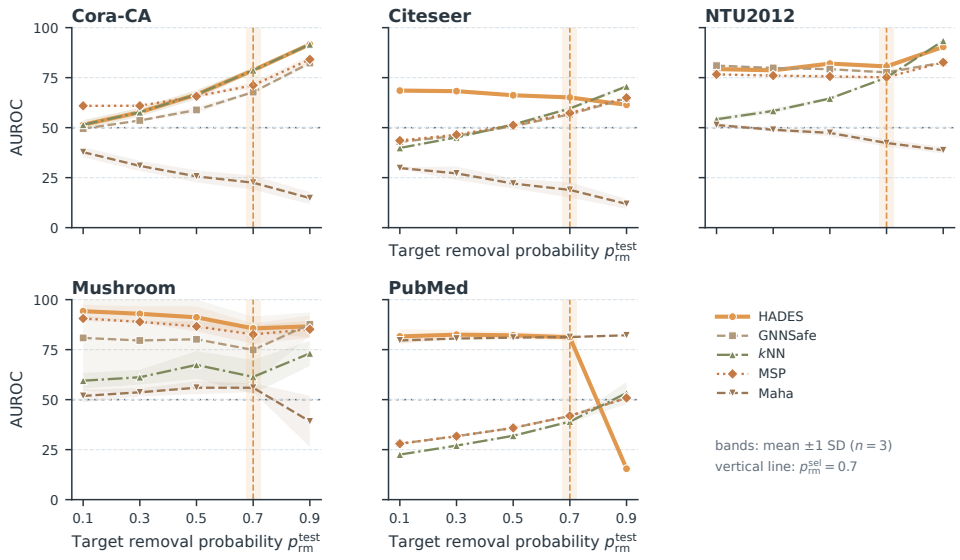


Figure 4: **Robustness to structural-shift severity** using the HGNN backbone. Curves show three-run mean AUROC with ± 1 standard-deviation bands. The detector is selected once at $p_{\mathrm{rm}}^{\mathrm{sel}} = 0.7$ and frozen as $p_{\mathrm{rm}}^{\mathrm{test}}$ varies. The dotted line marks random AUROC.

4.4 Hyperparameter Sensitivity

We next examine whether the discrepancy remains stable beyond a narrowly tuned setting. Figure 3 tests this dependence by varying K', τ, K, and α around their fixed values while holding the backbone and remaining settings fixed. The mean changes by at most 0.5 points on Cora-CA, 0.3 on Citeseer, and 0.01 on NTU2012 across every sweep. Mushroom is less stable. Its mean spans 9.5 points over τ, 17.6 over K, and 17.3 over α, with substantial variation across runs. The close full and reduced pool results on Mushroom show that another candidate can compensate when the discrepancy is sensitive to its parameters.

4.5 Robustness Across Structural Shifts

We test severity transfer in Figure 4 by selecting once at $p_{\mathrm{rm}}^{\mathrm{sel}} = 0.7$ and evaluating independent target copies from $p_{\mathrm{rm}}^{\mathrm{test}} = 0.1$ to 0.9 without reselection. The frozen choice ranks first or second among five detectors in 23 of 25 settings and rises from 51.4 to 91.7 AUROC on Cora-CA. Yet PubMed remains near 81 through $p_{\mathrm{rm}}^{\mathrm{test}} = 0.7$ before dropping to 15.5 at 0.9, exposing an extreme shift beyond the selected signal's coverage. Thus, selection transfers broadly, but not universally. Appendix B.2 evaluates independent incidence realizations.

5 Conclusion

HADES reframes structural OOD detection on hypergraphs as a joint problem of relational reference construction and detector selection. Its prototype anchored discrepancy and shift aligned self validation achieve the best average detector rank on all four evaluated backbones across 12 hypergraphs, showing that reliability need not rest on one universal score. The resulting view couples a clean relational reference with score selection guided by validation evidence. It opens a path toward broader shift models, uncertainty calibrated selection, and abstention based on validation evidence.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist literature retrieval and discovery, and support research ideation and execution. We have reviewed all AI assisted work and checked it against the original sources, method, code, and experimental records. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

HADES is evaluated on existing public benchmark releases and does not involve new data collection or human-subject studies. It provides a diagnostic signal rather than a safety guarantee: missed OOD nodes and false alarms remain possible, especially when the deployment shift differs from the self-validation family. Any application involving data about individuals requires a separate assessment of privacy, bias, fairness, and potential downstream harms in its intended setting. The authors are not aware of other ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the task, score, selection rule, and evaluation protocol. Appendix A.2 documents the fixed hyperparameters, detector candidates, implementation steps, and computational resources. Appendix A.1 describes the datasets, preprocessing, and splits. Appendix C.4 reports the complete per-dataset results. The supplementary materials provide a standalone implementation of HADES for backbone training, self validation, and target scoring.

REFERENCES

Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon M. Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018. doi: 10.1073/pnas.1800683115.

Derun Cai, Moxian Song, Chenxi Sun, Baofeng Zhang, Shenda Hong, and Hongyan Li. Hypergraph structure learning for hypergraph neural networks. In *International Joint Conference on Artificial Intelligence*, pp. 1923–1929, 2022. doi: 10.24963/ijcai.2022/267.

Tingyi Cai, Yunliang Jiang, Ming Li, Changqin Huang, Yi Wang, and Qionghao Huang. ML-GOOD: Towards multi-label graph out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):15650–15658, 2025. doi: 10.1609/aaai.v39i15.33718.

Tingyi Cai, Yunliang Jiang, Ming Li, Changqin Huang, Yujie Fang, Chengling Gao, and Zhonglong Zheng. HyperGOOD: Towards out-of-distribution detection in hypergraphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):19817–19825, 2026. doi: 10.1609/aaai.v40i24.39064.

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3D model retrieval. *Computer Graphics Forum*, 22(3):223–232, 2003. doi: 10.1111/1467-8659.00669.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Philip S. Chodrow, Nate Veldt, and Austin R. Benson. Generative hypergraph clustering: From blockmodels to modularity. *Science Advances*, 7(28):eabh1303, 2021. doi: 10.1126/sciadv.abh1303.

Qionghai Dai and Yue Gao. The DeepHypergraph library. In *Hypergraph Computation*, Artificial Intelligence: Foundations, Theory, and Algorithms, pp. 237–240. Springer, Singapore, 2023. doi: 10.1007/978-981-99-0185-2_12.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558.

Richard Forsyth. Zoo. UCI Machine Learning Repository, 1990. URL https://doi.org/10.24432/C5R59V.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Consistency of spectral hypergraph partitioning under planted partition model. *The Annals of Statistics*, 45(1):289–315, 2017. doi: 10.1214/16-AOS1453.

Mononito Goswami, Cristian Challu, Laurent Callot, Lenon Minorics, and Andrey Kan. Unsupervised model selection for time series anomaly detection. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gOZ_pKANaPW.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkxzx0NtDB.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, volume 162, pp. 8759–8773, 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html.

Yue Hou, Ruomei Liu, Yingke Su, Junran Wu, and Ke Xu. Graph out-of-distribution detection via test-time calibration with dual dynamic dictionaries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21797–21804, 2026. doi: 10.1609/aaai.v40i26.39331.

Jing Huang and Jie Yang. UniGNN: A unified framework for graph and hypergraph neural networks. In *International Joint Conference on Artificial Intelligence*, pp. 2563–2569, 2021. doi: 10.24963/ijcai.2021/353.

Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *International Joint Conference on Artificial Intelligence*, pp. 2635–2641, 2019. doi: 10.24963/ijcai.2019/366.

Ken Lang. NewsWeeder: Learning to filter netnews. In *International Conference on Machine Learning*, pp. 331–339, 1995. doi: 10.1016/B978-1-55860-377-6.50048-7.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Fan Li, Xiaoyang Wang, Wenjie Zhang, Ying Zhang, and Xuemin Lin. DHG-Bench: A comprehensive benchmark for deep hypergraph learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lhsb1ChUDF.

Ming Li, Yujie Fang, Yi Wang, Han Feng, Yongchun Gu, Lu Bai, and Pietro Liò. Deep hypergraph neural networks with tight framelets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18385–18392, 2025. doi: 10.1609/aaai.v39i17.34023.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the limits of softmax-based out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023. doi: 10.1109/CVPR52729.2023.02293.

Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 1686–1695, 2023. doi: 10.1109/ICCV51070.2023.00162.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-OOD detection. *arXiv*, 2021. doi: 10.48550/arXiv.2106.09022.

Chuancheng Song, Hanyang Shen, Yan Dong, Xixun Lin, Yanmin Shang, and Yanan Cao. Breaking one-size-fits-all: Revisiting out-of-distribution detection on graphs under diverse distribution shifts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25527–25535, 2026. doi: 10.1609/aaai.v40i30.39748.

Li Sun, Lanxu Yang, Jiayu Tian, Bowen Fang, Xiaoyan Yu, Junda Ye, Peng Tang, Hao Peng, and Philip S. Yu. Learning to explore: Policy-guided outlier synthesis for graph out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(30):25709–25717, 2026. doi: 10.1609/aaai.v40i30.39768.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, volume 162, pp. 20827–20840, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Bohan Tang, Zexi Liu, Keyue Jiang, Siheng Chen, and Xiaowen Dong. Training-free message passing for learning on hypergraphs. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4AuyYxt7A2.

Mushroom. UCI Machine Learning Repository, 1981. URL https://doi.org/10.24432/C5959T.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4911–4920, 2022. doi: 10.1109/CVPR52688.2022.00487.

Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=RiTjKoscnNd.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zoz7Ze4STUL.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920, 2015. doi: 10.1109/CVPR.2015.7298801.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1efa39bcaec6f3900149160693694536-Abstract.html.

Shenzhi Yang, Bin Liang, An Liu, Lin Gui, Xingkai Yao, and Xiaofang Zhang. Bounded and uniform energy-based out-of-distribution detection for graphs. In *International Conference on Machine Learning*, volume 235, pp. 56216–56234, 2024. URL https://proceedings.mlr.press/v235/yang24n.html.

Jiying Zhang, Yuzhao Chen, Xi Xiao, Runiu Lu, and Shu-Tao Xia. Learnable hypergraph laplacian for hypergraph learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4503–4507, 2022. doi: 10.1109/ICASSP43922.2022.9747687.

Yanan Zhao, Yuelong Li, Haichuan Zhang, Vishal Monga, and Yonina C. Eldar. Deep, convergent, unrolled half-quadratic splitting for image deconvolution. *IEEE Transactions on Computational Imaging*, 10:574–588, 2024. doi: 10.1109/TCI.2024.3377132.

Yanan Zhao, Xingchao Jian, Feng Ji, Wee Peng Tay, and Antonio Ortega. Generalized graph signal reconstruction via the uncertainty principle. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889406.

Yue Zhao, Ryan A. Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4489–4502, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/23c894276a2c5a16470e6a31f4618d73-Abstract.html.

SUPPLEMENTARY OVERVIEW

The supplement addresses four questions that connect reproducibility, mechanism, extended evidence, and research context to the main paper.

- **Can the evaluation be reproduced?** Appendix A specifies the datasets, perturbation protocol, fixed settings, detector scope, and the two model-level reproductions.
- **Why and when should HADES work?** Appendix B gives the formal conditions behind both components, then tests selector behavior under resampled and natural incidence shifts.
- **Do the conclusions persist beyond the main table?** Appendix C reports additional metrics, feature-shift and ID diagnostics, aggregate views, and the complete per-backbone results.
- **How does the method relate to adjacent work?** Appendix D expands the comparison with post-hoc detection, relational confidence, and learned-incidence methods.

A EXPERIMENTAL PROTOCOL AND REPRODUCIBILITY

A.1 DATASETS AND SPLITS

Table 3 lists the 12 hypergraph benchmarks. The co-authorship and co-citation hypergraphs follow the DHG-Bench/HyperGCN releases (Yadati et al., 2019; Dai & Gao, 2023; Li et al., 2026). The shape and categorical collections follow the processed AllSet/HyperGOOD releases (Chien et al., 2022; Cai et al., 2026). Zoo and Mushroom originate from the UCI repository (Forsyth, 1990; UCI, 1981). 20News derives from the 20 Newsgroups collection (Lang, 1995). The shape benchmarks derive from NTU (Chen et al., 2003) and ModelNet40 (Wu et al., 2015). Senate and House are legislative committee hypergraphs (Benson et al., 2018; Chodrow et al., 2021).

For Cora-CA, Cora-CC, Citeseer, PubMed, and DBLP, we retain the DHG training masks and partition the remaining nodes into a 500-node validation set and a disjoint test set. For the other seven datasets, we generate one 50/25/25 train/validation/test split with seed 42, following the AllSet protocol. The released masks and generated splits are fixed across all methods and runs. Structural OOD evaluation pairs the same test nodes under the clean incidence and a perturbed copy. Node features, labels, and split indices remain fixed. The suite ranges from 101 to 41,302 nodes, spanning nearly three orders of magnitude.

Senate and House have no native node attributes. Following the HyperGOOD-compatible benchmark, we use its released 100-dimensional processed features unchanged. These combine a one-hot encoding of the class label with Gaussian noise. We mark both datasets in Table 3 and report their clean ID accuracies in Table 10, since this construction supplies class information directly to the backbone. We retain these datasets for benchmark comparability.

Table 3: **Benchmark statistics.** N nodes, M hyperedges, F feature dimensions, C classes, and the ID train/validation/test split. † denotes a retained DHG training mask with 500 validation nodes. ‡ denotes a 50/25/25 split. Asterisks mark label-derived released features.

Dataset	**Domain**	N	M	F	C	**Train / Val / Test**
Cora-CA†	co-authorship	2,708	970	1,433	7	140 / 500 / 2,068
Cora-CC†	co-citation	2,708	1,483	1,433	7	140 / 500 / 2,068
Cite.†	co-citation	3,312	1,004	3,703	6	138 / 500 / 2,674
PubM.†	co-citation	19,717	7,531	500	3	78 / 500 / 19,139
NTU‡	3D shape	2,012	1,635	100	67	1,006 / 503 / 503
M.Net‡	3D shape	12,311	10,630	100	40	6,155 / 3,077 / 3,079
Mush.‡	categorical	8,124	287	22	2	4,062 / 2,031 / 2,031
Sen.*‡	committees	282	301	100	2	141 / 70 / 71
House*‡	committees	1,290	335	100	2	645 / 322 / 323
DBLP†	co-authorship	41,302	20,865	1,425	6	1,740 / 500 / 39,062
Zoo‡	categorical	101	40	16	7	50 / 25 / 26
20NG‡	text	16,242	100	100	4	8,121 / 4,060 / 4,061

A.2 HADES IMPLEMENTATION

Backbone training. We select HGNN, UniGCNII, AllSetTransformer, and ED-HNN because they represent spectral propagation, residual message passing, set based aggregation, and equivariant diffusion, respectively (Feng et al., 2019; Huang & Yang, 2021; Chien et al., 2022; Wang et al., 2023). This coverage lets us assess whether detector behavior transfers across distinct HNN inductive biases rather than a single encoder family. For each dataset and architecture, the backbone is trained on the clean hypergraph with cross-entropy over the released training split, and its checkpoint is selected by clean validation accuracy. All detectors that leave the training objective unchanged, including HADES, reuse this checkpoint. NODESAFE is trained separately with its bounded-energy regularizer. Every reported detector uses three independent runs indexed by global master seeds 42, 123, and 456. Model initialization and training use the corresponding master seed.

Structural perturbation. To isolate whether a detector recognizes relational change, we construct a shifted copy in which the nodes, features, labels, and split indices remain fixed and only the incidence changes. Removing hyperedges represents the dissolution of familiar groups, while adding new ones introduces previously unseen memberships for the same nodes. For every original hyperedge, we draw an independent Bernoulli keep variable with probability $1 - p_{\mathrm{rm}}$. If M_{keep} edges survive, we add $\lfloor p_{\mathrm{add}} M_{\mathrm{keep}} \rfloor$ random hyperedges. Each new edge size is sampled from the surviving edge-size multiset, and its incident nodes are sampled uniformly without replacement. The main setting is $p_{\mathrm{rm}} = 0.7$, $p_{\mathrm{add}} = 0.5$. Thus the number added is half the number surviving, not the number removed. The target structural copy is sampled independently in each run. Its random stream is derived deterministically from that run's master seed and is disjoint from model training and HADES self-validation. All detectors compared within a cell receive the same target copy.

Confidence and propagation. The discrepancy should reflect relational support rather than branch specific confidence or propagation settings. We therefore initialize both views with the same negative free energy confidence, $T \log \sum_c \exp(Z_c/T)$ with $T = 1$, and apply the same SEP dynamics with $K = 2$ and $\alpha = 0.5$. These propagation settings also match our GNNSafe and HyperGOOD implementations (Wu et al., 2023; Cai et al., 2026). Throughout the paper, larger confidence scores indicate ID.

Prototype anchor. The anchor must retain clean representation geometry while still locating every node in the currently scored hypergraph. On the first clean call, ten iterations of spherical k-means fit $K' = \max(8, \min(64, 4C))$ centers to normalized train-node embeddings, and the centers are cached for the run. Each subsequent call reassigns the current embeddings to these fixed centers with temperature $\tau = 0.5$. Thus the coordinate system remains tied to the clean regime, while the soft memberships can respond to the current node representations without using current hyperedge memberships to construct the anchor. Before comparison, each propagated vector is standardized on the fixed train indices. the standard-deviation denominator is clamped below by 10^{-6}.

Candidate pool and self-validation. Because incidence shift may appear through relational disagreement, confidence, or feature geometry, the fixed pool contains s_{disc}, GNNSafe, Energy, MSP, ODIN, kNN, Mahalanobis, and ViM. Self-validation estimates which signal transfers to the anticipated shift by averaging validation-node proxy AUROC over $R = 3$ perturbation copies. These copies come from a second deterministic stream derived from the run's master seed and remain independent of the target copy. The resulting candidate is cached before the target hypergraph is scored, so target scores and OOD labels cannot influence the choice.

The proxy is evaluated on held-out validation nodes rather than the training nodes used to build feature-space references. Otherwise, kNN and Mahalanobis can benefit from self-matches and yield an inflated reliability estimate. On Cora-CA with UniGCNII, for example, the training-node kNN proxy is 90.9 even though its target structural OOD AUROC is about 63. Using the same held-out indices for every candidate makes the proxy measure transfer beyond each candidate's reference set.

Algorithm 1 summarizes the resulting chronology: fit the clean anchor, generate surrogates, rank the candidates on validation nodes, freeze one detector, and only then score the target hypergraph.

Algorithm 1 HADES self validated selection

Input: frozen backbone f, clean hypergraph $(\mathbf{X}, \mathbf{H})$, train and validation indices, candidate pool $\mathcal{D}$, perturbation generator Q, number of copies R

Output: cached detector $d^\star$ and target scores $d^\star(\mathbf{X}, \widetilde{\mathbf{H}})$

Fit and cache prototype centers on clean train embeddings
for $r = 1, \ldots, R$ **do**
    $\mathbf{H}_{\mathrm{sur}}^{(r)} \sim Q(\mathbf{H})$
end for
for $d \in \mathcal{D}$ **do**
    Evaluate d on all incidences, using Eq. (5) when $d = s_{\mathrm{disc}}$
    $a_d \leftarrow R^{-1} \sum_r \mathrm{AUROC}_{\mathrm{val}}\big(d(\mathbf{X}, \mathbf{H}), d(\mathbf{X}, \mathbf{H}_{\mathrm{sur}}^{(r)})\big)$
end for
$d^\star \leftarrow \arg\max_d a_d$
return cached $d^\star$ and evaluate $d^\star(\mathbf{X}, \widetilde{\mathbf{H}})$ at deployment

Compute environment. Scalability follows from representing the clean reference as an $N \times K'$ soft incidence rather than a dense node by node matrix. Fitting and assigning the prototypes costs $O(NK'd)$, prototype SEP costs $O(KNK')$, and current-incidence SEP costs $O(K\,\mathrm{nnz}(\widetilde{\mathbf{H}}))$, with $O(NK')$ memory for the prototype incidence. As a one-time pre-deployment cost, self-validation evaluates every candidate on the clean graph and on R surrogate copies. On one RTX 5060 Ti with 16 GB memory, this implementation completed all 12 benchmarks without a dataset-specific fallback.

A.3 Compared Detectors

We compare 13 fixed detectors with HADES, plus HyperGOOD and UniGOD-H as separate model-level references.

- **Confidence and logit scores:** MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Energy (Liu et al., 2020), MaxLogit and KL-Matching (Hendrycks et al., 2022), and GEN (Liu et al., 2023).
- **Feature-space scores:** Mahalanobis (Lee et al., 2018), Relative-Mahalanobis (Ren et al., 2021), kNN (Sun et al., 2022), ViM (Wang et al., 2022), and NNGuide (Park et al., 2023).
- **Relational methods:** GNNSafe (Wu et al., 2023), NODESAFE (Yang et al., 2024), HyperGOOD (Cai et al., 2026), and UniGOD-H, our hypergraph adaptation of UniGOD (Song et al., 2026).

The first 11 and GNNSafe are applied to the same standard-trained backbone used by HADES. NODESAFE changes the training objective and is therefore retrained for each architecture. HyperGOOD and UniGOD-H use their own reproduced model backbones. GNNSafe, NODESAFE, and our HyperGOOD reproduction use the symmetric normalized hypergraph operator in Eq. (3).

Scope of recent graph OOD methods. We do not place ML-GOOD, BaCa, or PGOS in the main comparison because their published tasks do not match our single-label, node-level structural OOD protocol. ML-GOOD is designed for multi-label node classification and aggregates label-specific energies (Cai et al., 2025). BaCa and PGOS operate at graph level, treating each graph as one sample. BaCa updates dynamic dictionaries from the target test batch, whereas PGOS synthesizes complete pseudo-OOD graphs (Hou et al., 2026; Sun et al., 2026). UniGOD is node-level and includes a structure-manipulation benchmark, making it the closest of these recent graph methods to our setting (Song et al., 2026). However, UniGOD trains a curvature adaptive GNN with neural SDE dynamics rather than providing a post hoc score for the four frozen HNN backbones. We therefore report it as the separate model level reproduction described next rather than including it in the candidate pool or four backbone detector ranking.

A.4 UniGOD-H Reproduction

UniGOD-H adapts UniGOD (Song et al., 2026) to hypergraphs by replacing its graph-neighborhood aggregation with the symmetric normalized hypergraph propagation in Eq. (3). The three curvature-constrained spaces, geometry and dynamics coupling objective, trajectory uncertainty, and multiplicative distance and uncertainty score remain unchanged. The resulting method is a complete trained model and is evaluated as a separate model-level reference rather than a post-hoc candidate available to HADES.

The evaluation follows the shared three-run protocol in Section 4.1, the splits in Appendix A.1, and the main perturbation setting $p_{\mathrm{rm}} = 0.7$ and $p_{\mathrm{add}} = 0.5$. Each run trains for at most 300 epochs with early-stopping patience 60, selects the checkpoint by clean validation accuracy, and applies `ReduceLROnPlateau` with patience 20 and factor 0.5. UniGOD-H uses two layers, hidden dimension 64, dropout 0.5, and Adam with learning rate 10^{-3} and weight decay 5×10^{-4}. We set the SDE parameters to $t = 4$ and $\Delta t = 0.5$ within the effective range recommended by UniGOD. Parameters not specified in the original paper are fixed before target evaluation: $\alpha = 1$, $\beta = 0.5$, unit weights on the three losses, and 2 and 8 Monte Carlo samples for training and evaluation, respectively. No target OOD data are used for tuning, and all results use ID-positive AUROC, AUPR, and FPR95.

Table 4: **Full UniGOD-H results** under the HADES structural OOD protocol ($\times 100$). Per-dataset entries are mean$\pm$standard deviation over three runs, and the final row averages the dataset means. Higher is better for AUROC, AUPR, and ID accuracy. Lower is better for FPR95.

Dataset	AUROC ↑	AUPR ↑	FPR95 ↓	ID Acc. ↑
Cora-CA	68.43±2.32	63.43±1.74	82.66±3.12	27.10±0.14
Cora-CC	53.85±0.28	56.03±0.86	95.58±0.18	13.23±4.04
Cite.	44.26±0.70	46.35±0.35	96.77±0.64	12.79±0.00
PubM.	57.90±0.48	54.74±0.31	90.10±0.43	35.14±0.00
NTU	55.27±2.66	53.43±0.79	87.08±6.87	78.93±1.65
M.Net	56.39±0.99	54.67±0.41	88.38±5.31	94.20±0.31
Mush.	49.10±3.03	48.67±1.51	95.21±1.77	84.77±1.42
Sen.	50.64±0.61	53.03±0.82	93.43±2.89	100.00±0.00
House	52.43±0.30	51.78±0.35	95.05±0.88	100.00±0.00
DBLP	69.83±0.60	64.69±0.55	61.42±0.68	29.59±0.89
Zoo	48.42±1.46	50.61±3.13	98.72±1.81	82.05±1.81
20NG	54.53±4.15	53.03±2.33	92.35±3.72	59.05±1.52
Average	55.09	54.21	89.73	59.74

Across the 12 datasets, UniGOD-H obtains average AUROC 55.1, AUPR 54.2, FPR95 89.7, and ID accuracy 59.7. HADES using the HGNN backbone is higher in AUROC on 11 datasets, while UniGOD-H leads only on Senate. UniGOD-H is below random AUROC on Citeseer, Mushroom, and Zoo. Under this controlled adaptation, replacing graph aggregation with hypergraph propagation alone does not yield a consistently reliable higher-order structural OOD signal.

A.5 HyperGOOD Reproduction

HyperGOOD (Cai et al., 2026) was published for feature and label perturbations under fixed incidence. To obtain a comparable structural reference, we port its framelet backbone and structure-aware energy propagation into our evaluation harness. The reproduction follows the shared splits, three global master seeds, paired target copies, and ID-positive metrics in Section 4.1. Checkpoints are selected only by clean validation accuracy. Its energy propagation uses the symmetric normalized hypergraph operator and the settings specified in Appendix A.2. The values in Table 1 are therefore results under our structural protocol rather than numbers copied from the published feature/label experiments.

Table 5: **Full HyperGOOD reproduction** under the HADES structural OOD protocol ($\times 100$). Per-dataset entries are mean$\pm$standard deviation over three runs. The final row averages the completed-dataset means. N/A denotes runs that exceeded the stated resource budget.

Dataset	AUROC ↑	AUPR ↑	FPR95 ↓	ID Acc. ↑
Cora-CA	72.86±0.63	69.85±0.26	83.82±1.84	65.62±0.12
Cora-CC	54.94±0.93	55.35±0.84	93.87±0.15	52.58±4.04
Cite.	57.16±0.43	57.32±0.36	93.59±0.25	43.03±1.59
PubM.	42.18±0.46	43.88±0.31	95.92±0.07	36.40±1.78
NTU	77.29±0.81	77.33±1.09	69.45±3.79	81.44±0.34
M.Net	75.14±0.33	74.87±0.41	70.08±0.48	94.21±0.31
Mush.	N/A	N/A	N/A	N/A
Sen.	44.24±2.31	51.83±1.79	97.65±0.66	100.00±0.00
House	47.84±0.95	50.38±0.68	92.67±1.54	100.00±0.00
DBLP	77.47±0.21	72.80±0.33	55.71±0.48	85.92±0.07
Zoo	67.55±10.98	71.47±11.61	71.79±10.09	73.08±5.44
20NG	N/A	N/A	N/A	N/A
Average over completed datasets	61.67	62.51	82.46	73.23

We apply a fixed resource screen before training and skip framelet construction when $\sum_{e\in\mathcal{E}} |e|^2 > 1.5\times 10^7$, because the resulting clique-induced operator products would exceed the available 64 GB host memory. Mushroom and 20News cross this threshold and are marked N/A. The screen records the budget of our reproduction rather than a general limitation of HyperGOOD. Across the ten completed datasets, HADES using the HGNN backbone reaches 69.7 AUROC, 72.8 AUPR, and 82.5 FPR95, compared with 61.7, 62.5, and 82.5 for the reproduced HyperGOOD model. HADES leads HyperGOOD in AUROC on eight datasets, and the average improvement is concentrated in AUROC and AUPR while mean FPR95 is unchanged.

B Mechanism and Shift Reliability

B.1 Mechanistic Conditions and Selection Bound

This section formalizes the conditions behind Proposition 1 and the selection-regret statement following Eq. (7).

For an evaluated incidence $\mathbf{A} \in \{\mathbf{H}, \widetilde{\mathbf{H}}\}$, let $\mathbf{q}_{\mathbf{A}}$ denote the confidence vector obtained from the frozen backbone on $(\mathbf{X}, \mathbf{A})$, and let $\widehat{\mathbf{H}}_{\mathbf{A}}$ be the prototype incidence formed by assigning its embeddings to the fixed clean centers. Using the train-index standardization z_{tr} from Section 3.1, define

$$a_{\mathbf{A}}(v) = z_{\mathrm{tr}}(\mathrm{SEP}_{\mathbf{A}}(\mathbf{q}_{\mathbf{A}}))_v\,, \qquad b_{\mathbf{A}}(v) = z_{\mathrm{tr}}\Big(\mathrm{SEP}_{\widehat{\mathbf{H}}_{\mathbf{A}}}(\mathbf{q}_{\mathbf{A}})\Big)_v\,,$$

so that $D_{\mathbf{A}}(v) = |a_{\mathbf{A}}(v) - b_{\mathbf{A}}(v)|$.

Assumption 1 (Standardized community signal)**.** *Each node belongs to a latent community $c(v)$. There exist community values m_c and $\beta \geq 0$ such that, on the clean hypergraph, both standardized streams are within β of $m_{c(v)}$. Under a structural shift, the prototype-anchored stream remains within $\beta + \eta_v$ of $m_{c(v)}$ for some $\eta_v \geq 0$, while the current-incidence stream is within β of*

$$m_{c(v)} + \rho_v \sum_{c\neq c(v)} \pi_{vc}\big(m_c - m_{c(v)}\big).$$

Here $\rho_v \in [0,1]$ is the cross-community propagation mass, while $\pi_{vc} \geq 0$ and $\sum_{c\neq c(v)} \pi_{vc} = 1$ describe its community composition. Define the residual contrast $\Gamma_v = \Big|\sum_{c\neq c(v)} \pi_{vc}(m_c - m_{c(v)})\Big|$.

The assumption is motivated by planted-partition analyses of hypergraph spectral operators (Ghoshdastidar & Dukkipati, 2017). The term η_v bounds drift caused by assigning shifted embeddings to fixed clean centers, while Γ_v records the community contrast remaining after oppositely signed changes cancel.

Proof of Proposition 1. For the clean hypergraph, both $a_{\mathbf{H}}(v)$ and $b_{\mathbf{H}}(v)$ are within β of the same $m_{c(v)}$, so the triangle inequality gives $D_{\mathbf{H}}(v) \leq 2\beta$. Under the shift, the centers of the two assumed intervals differ in magnitude by $\rho_v \Gamma_v$. Subtracting their maximum deviations, β for the current-incidence stream and $\beta + \eta_v$ for the prototype-anchored stream, gives the lower bound. Comparing it with the clean upper bound yields the separation condition. □

Proposition 1 requires structural contamination to reach the node ($\rho_v > 0$), community-level confidence changes not to cancel ($\Gamma_v > 0$), and prototype-anchor drift to remain controlled. When any condition fails, s_{disc} need not be useful. The selector can instead retain a complementary candidate.

Proposition 2 (Surrogate-selection regret)**.** *Let $\widehat{R}(d)$ and $d^\star$ be defined by Eq. (7), and let $R(d)$ be the target AUROC. If $\sup_{d\in\mathcal{D}} |\widehat{R}(d) - R(d)| \leq \varepsilon$, then*

$$R(d^\star) \geq \max_{d\in\mathcal{D}} R(d) - 2\varepsilon.$$

Proof of Proposition 2. Let $d_{\mathrm{opt}} = \arg\max_{d\in\mathcal{D}} R(d)$. Under the proposition's uniform bound,

$$R(d^\star) \geq \widehat{R}(d^\star) - \varepsilon \geq \widehat{R}(d_{\mathrm{opt}}) - \varepsilon \geq R(d_{\mathrm{opt}}) - 2\varepsilon.$$

The middle inequality is the definition of $d^\star$. □

Surrogate-error decomposition. Condition on the fixed validation split and trained models, and let $\bar{R}(d) = \mathbb{E}_Q[\widehat{R}(d)]$, where the expectation is over the surrogate generator Q. Define

$$\varepsilon_{\mathrm{mis}} = \sup_{d\in\mathcal{D}} |\bar{R}(d) - R(d)|, \qquad \varepsilon_{\mathrm{mc}} = \sup_{d\in\mathcal{D}} |\widehat{R}(d) - \bar{R}(d)|.$$

The first term captures systematic mismatch between proxy and target performance, including differences between the surrogate and target shift families and between validation and target populations. The second captures variability from finitely many surrogate copies. The triangle inequality gives $\sup_{d\in\mathcal{D}} |\widehat{R}(d) - R(d)| \leq \varepsilon_{\mathrm{mis}} + \varepsilon_{\mathrm{mc}}$. Proposition 2 therefore yields regret at most $2(\varepsilon_{\mathrm{mis}} + \varepsilon_{\mathrm{mc}})$. This decomposition distinguishes systematic mismatch from finite-copy variability. The regret result itself is deterministic and does not require a probability model for either term.

B.2 Robustness to OOD-Copy Sampling

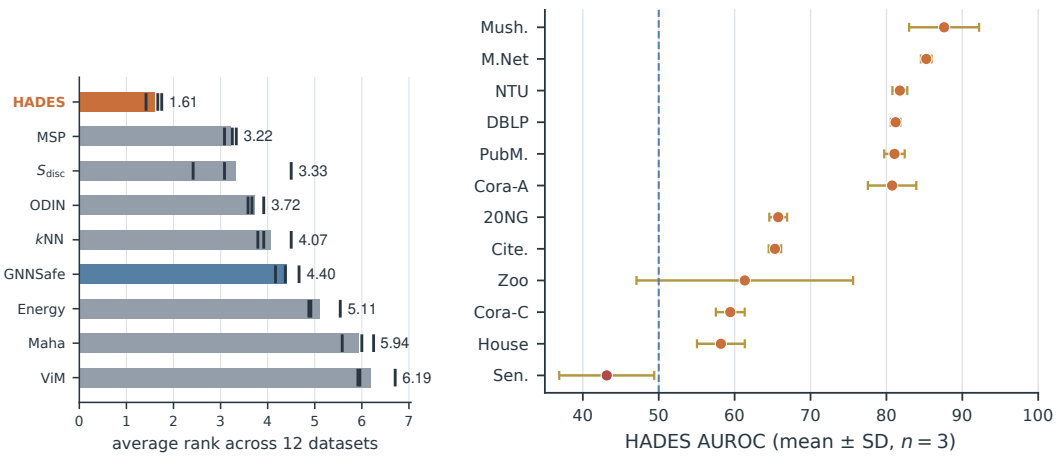
Table 6 and Figure 5 evaluate sensitivity to the sampled target perturbation by repeating the HGNN evaluation over three independently sampled structural OOD copies. Baseline and HADES scores use the same copy within each run. HADES remains first among the eight candidates on every copy, with average rank ranging from 1.42 to 1.75. Across datasets, its AUROC standard deviation over the three copies averages 3.3. The 14.3 maximum occurs on Zoo, whose test split has only 26 nodes.

Table 6: **Sampling robustness** using the HGNN backbone. Rank is among the eight HADES candidate signals over 12 datasets. Within 1 counts datasets where HADES is no more than one AUROC point below the strongest candidate.

Run	HADES average rank ↓	HADES wins / 12	within 1 / 12
1	1.42	9	9
2	1.75	8	9
3	1.67	9	10

B.3 Natural Incidence Transfer

The edge removal benchmark controls structural shift while preserving the source hypergraph's hyperedge formation mechanism. We therefore use natural transfer to test whether a detector selected from these surrogates remains reliable when the relation itself changes. Cora provides a controlled test because its coauthorship and cocitation views share the same 2,708 nodes, features, labels, and splits. Their incidences contain 970 and 1,483 hyperedges, respectively, and encode different relations. In each direction, we train the HGNN and fit the prototypes on the source view, then select



(a) Rank across three runs **(b) Per-dataset run variation**

Figure 5: **Robustness to OOD-copy sampling.** Bars in (a) show each candidate's mean rank, with ticks for the three runs. Points and error bars in (b) show HADES mean AUROC and standard deviation. The dashed line marks random AUROC.

the detector from edge removal copies generated from that view. We freeze all components, treating source view test nodes as ID and the same nodes under the target incidence as OOD. This isolates relational change from node identity and attributes.

Table 7: **Natural Cora structure shift** (AUROC↑, mean±standard deviation over three runs, HGNN). The node set and features are fixed while train and target incidences use coauthorship or cocitation. Bold marks the best result in each direction.

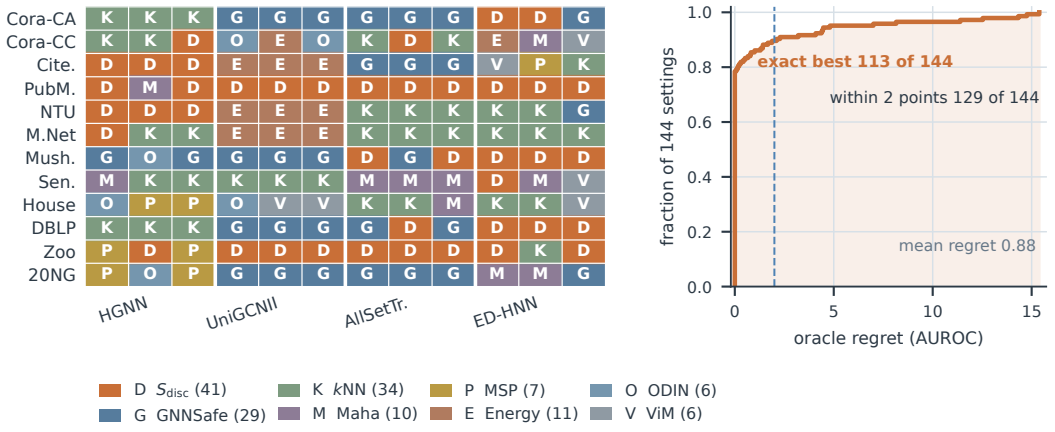
Detector	coauthorship → cocitation	cocitation → coauthorship
HADES	**78.2±1.4**	37.7±12.9
GNNSafe	74.8±0.0	25.8±0.4
Energy	51.6±17.2	62.1±4.0
MSP	71.3±0.5	31.8±0.3
ODIN	74.9±0.1	27.3±0.2
Maha	22.5±0.5	60.6±2.0
kNN	**78.2±1.4**	30.4±1.3
ViM	20.7±3.6	**70.4±1.8**

Table 7 reveals strong directional asymmetry in surrogate transfer. From co-authorship to co-citation, self-validation selects kNN in all three runs. HADES therefore matches the strongest candidate at 78.2 ± 1.4 AUROC and exceeds the next-best ODIN by 3.3 points. In the reverse direction, it selects kNN twice and s_{disc} once, reaching 37.7 ± 12.9, while ViM reaches 70.4 ± 1.8. Thus, perturbations constructed from the co-citation source do not anticipate which signal is reliable after transfer to co-authorship. The two directions show that transfer follows whether the surrogate preserves the target candidate ordering. Selection adapts when the ordering is preserved and weakens when a change in relation mechanism alters that ordering.

B.4 Selection Diagnostics

We next examine why the adaptive pool improves detection. The analysis asks whether synthetic validation identifies a strong candidate and whether s_{disc} contributes a useful structural signal without dominating every setting.

Figure 6 evaluates the selector at the decision level. The selected candidate attains the highest target AUROC in 113 of 144 settings and comes within two AUROC points in 129. Its mean oracle regret is 0.88 AUROC points. Figure 7 then tests whether these decisions reflect consistent agreement between synthetic validation and target performance. Across every candidate and setting, the pooled



(a) Selected score by setting **(b) Oracle regret**

Figure 6: **Selections and regret across 144 combinations of dataset, backbone, and run.** *(a)* Selected candidate in each run and aggregate selection frequency. Within every backbone, columns follow seeds 42, 123, and 456. *(b)* Regret to an oracle selector that chooses the best candidate in the fixed pool using target labels. The oracle is diagnostic only and is never used by HADES.

Pearson correlation is $r = 0.95$. Because pooled correlation can be inflated by differences in setting difficulty, we also compute it within each setting. The mean is 0.86, the median is 0.98, and 141 of 144 correlations are positive. These correlations show that synthetic validation usually preserves the candidate ordering needed for selection. They also explain the asymmetric natural transfer result in Appendix B.3, where selection weakens when that ordering changes across relation mechanisms.

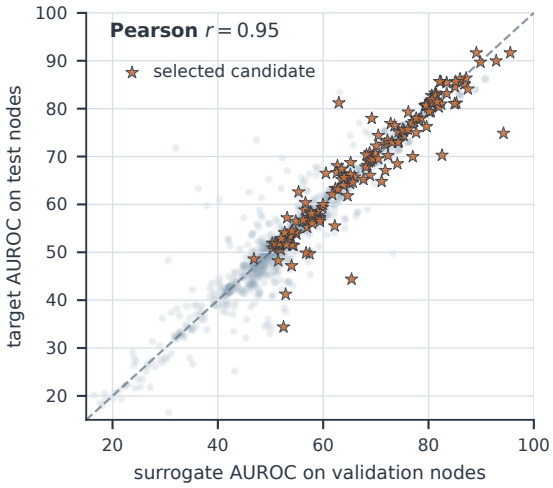
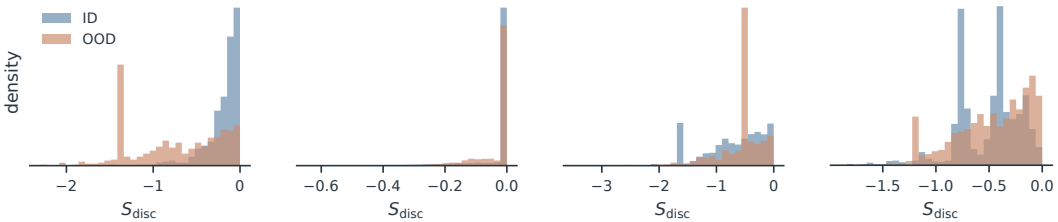


Figure 7: **Synthetic proxy and target AUROC** for all eight candidates across 144 settings. Stars mark selected candidates. The pooled Pearson correlation is 0.95. Correlations computed within each setting are summarized in the text.

Selection accuracy alone does not establish that s_{disc} adds a distinct structural signal. Figure 8 therefore isolates its score distributions. The score separates ID and OOD nodes on NTU2012 and PubMed, reaching 83 AUROC on both datasets. The distributions overlap on Cora-CA and Mushroom, where the score reaches 44 and 45. In those HGNN settings, the selector instead chooses kNN and GNNSafe. This contrast supports the intended division of labor. The discrepancy supplies a useful structural candidate, while the selector prevents it from being treated as a universal detector.



(a) NTU2012 s_{disc} AUROC 83 **(b) PubMed** s_{disc} AUROC 83 **(c) Cora-CA** s_{disc} AUROC 44 **(d) Mushroom** s_{disc} AUROC 45

Figure 8: **Discrepancy score distributions** with HGNN and master seed 42. ID denotes clean test nodes and OOD denotes the same nodes under the perturbed incidence.

C Complete and Complementary Evaluation

C.1 Backbone and Metric Transfer

Changing either the encoder or evaluation metric may alter detector rankings. Table 8 therefore tests whether the HGNN result transfers across the four evaluated backbones and all three detection metrics. HADES has the best average AUROC rank on every backbone, ranging from 1.62 to 2.67. Using HGNN, it reaches 72.5 mean AUPR and 77.9 mean FPR95. The strongest fixed baselines reach 64.4 AUPR with NNGuide and 78.4 FPR95 with GEN. Its FPR95 rank of 4.29 is less dominant than its AUROC rank of 2.08, locating the clearest transfer in AUROC.

Table 8: **Stability across backbones and metric coverage.** Left: average AUROC rank over 12 datasets among HADES and 13 fixed detectors (lower is better). Right: mean detection metrics using the HGNN backbone. Representative baselines are shown. Appendix C.4 reports all detectors and datasets.

	Average AUROC rank ↓					**HGNN mean**		
Detector	**HGNN**	**UniGCNII**	**AllSetTr.**	**ED-HNN**	**Overall**	**AUROC↑**	**AUPR↑**	**FPR95↓**
HADES	**2.08**	**2.25**	**1.62**	**2.67**	**2.16**	**70.9**	**72.5**	**77.9**
GNNSafe	8.00	7.04	5.54	6.29	6.72	61.4	60.0	80.9
NODESAFE	9.00	7.46	6.50	6.88	7.46	58.7	57.9	85.9
KLM	9.58	6.75	3.83	3.12	5.82	49.8	56.8	92.4
Energy	8.75	5.00	8.21	9.25	7.80	58.0	59.3	89.4
MSP	5.62	8.50	8.50	8.08	7.68	63.3	62.0	80.4
kNN	7.67	9.12	4.46	6.83	7.02	59.7	62.6	85.4
Maha	10.83	11.58	11.17	10.88	11.11	43.8	51.6	95.5

Table 9 extends Table 8 to the full detector set using the HGNN backbone. The released TF-HNN implementation caches its training free diffusion for repeated transductive inference (Tang et al., 2025). Under incidence replacement, this cache reuses the clean graph representation and makes ordinary clean and perturbed node scores identical. We therefore omit TF-HNN from comparative statistics and full performance tables. This leaves four backbones whose representations are recomputed for every supplied incidence.

C.2 Feature Perturbations and ID Accuracy

Feature perturbations and clean ID accuracy isolate two factors behind the structural results. The first tests whether detector ordering transfers to another shift family. The second measures the quality of the representation on which each detector relies.

Feature perturbations. Table 11 keeps incidence fixed and perturbs features. MSP has the highest mean AUROC at 55.3, while the leading detector changes across datasets. This reordering shows that structural and feature perturbations expose different score families.

Table 9: **All three OOD metrics using the HGNN backbone, full detector set**. Results are means over 12 datasets under structural OOD. Bold marks the best result. HADES is best on every metric and in average rank against all 13 detectors.

Detector	AUROC↑	rank	AUPR↑	rank	FPR95↓	rank
Hypergraph / graph-native						
HADES	**70.9**	**2.08**	**72.5**	**2.67**	**77.9**	**4.29**
GNNSafe	61.4	8.00	60.0	7.46	80.9	7.75
NODESAFE	58.7	9.00	57.9	9.21	85.9	9.17
Transferred generic						
Energy	58.0	8.75	59.3	8.00	89.4	7.54
NNGuide	63.2	5.08	64.4	4.50	82.8	7.04
MSP	63.3	5.62	62.0	6.33	80.4	6.04
ODIN	63.2	6.42	62.0	6.29	81.3	6.92
MaxLogit	63.1	5.33	62.5	5.38	83.4	7.79
GEN	64.4	5.46	62.8	5.83	78.4	5.75
KLM	49.8	9.58	56.8	9.33	92.4	7.50
kNN	59.7	7.67	62.6	7.33	85.4	7.42
Maha	43.8	10.83	51.6	11.17	95.5	7.88
RMD	40.1	10.58	50.6	10.33	99.1	10.88
ViM	40.3	10.58	48.0	11.17	96.7	9.04

Table 10: **In distribution node classification accuracy** (%, mean over three runs) of each backbone used by the detectors.

Dataset	HGNN	UniGCNII	AllSetTr.	ED-HNN
Cora-CA	63.3	72.0	68.5	23.1
Cora-CC	45.0	68.3	53.5	23.4
Cite.	38.0	60.3	48.9	29.1
PubM.	23.3	37.8	39.1	39.0
NTU	86.3	88.8	86.2	6.9
M.Net	91.8	97.7	96.5	18.1
Mush.	96.6	99.8	99.8	92.5
Sen.	50.7	100.0	99.1	98.1
House	61.9	99.8	100.0	99.5
DBLP	88.0	84.7	87.2	31.2
Zoo	57.7	88.5	74.4	47.4
20NG	81.8	82.2	82.1	80.5

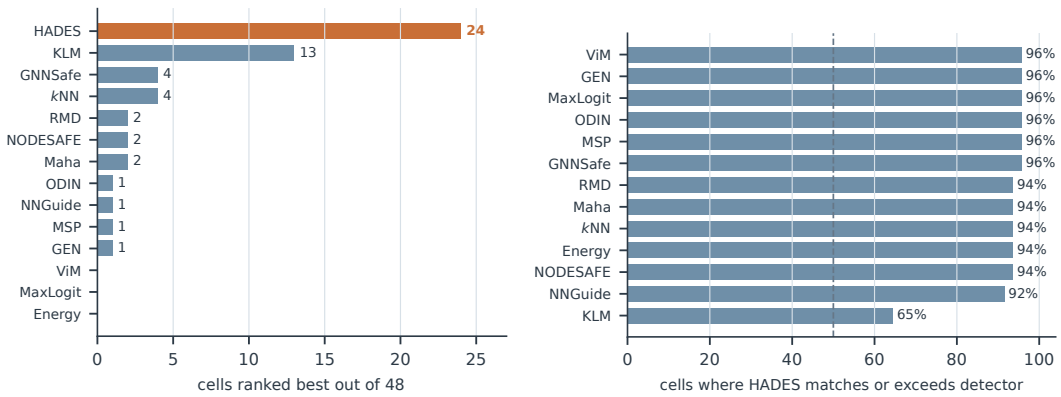
ID accuracy. Table 10 reports clean node classification accuracy. ED-HNN reaches only 6.9 on NTU and 18.1 on ModelNet, which helps explain why detectors built on this representation are less stable. The near perfect Senate and House values reflect the released label derived features described in Appendix A.1. Across the remaining cells, clean accuracy varies with both dataset and backbone. UniGCNII and AllSetTransformer are strongest on most noncitation benchmarks, while citation hypergraphs show a wider spread across architectures.

C.3 Additional Aggregate Views

Means across backbones can conceal whether an advantage is widespread or driven by a few cells. Figures 9 and 10 resolve the 48 dataset and backbone cells behind Table 8. HADES ranks first in 24 cells, compared with 13 for KLM, and matches or exceeds every fixed detector in at least 65 percent of cells. This coverage exceeds 91 percent for every detector except KLM. Its cellwise rank has mean 2.16 and median 1.75, placing most of its mass near the top while identifying KLM as the strongest specialized competitor.

Table 11: **Feature perturbation OOD using the HGNN backbone** ($\times 100$ AUROC, mean over three runs). Features are interpolated while incidence is fixed. Bold marks the best displayed result in each dataset. The changed ordering shows that structural and feature shifts favor different scores.

Detector	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**
GNNSafe	50.1	50.1	50.5	50.0	**56.4**	56.5	58.8	50.2	50.6	53.6	48.1	**55.7**	52.5
NODESAFE	51.2	50.0	50.0	50.0	56.2	55.9	51.7	50.3	50.5	50.5	48.2	51.6	51.4
Energy	50.9	**51.6**	**50.9**	50.0	56.3	**56.9**	56.3	50.3	51.6	56.2	48.8	53.7	52.8
MSP	**52.8**	50.8	50.5	50.0	55.4	56.4	**64.6**	**52.1**	**52.4**	**56.7**	**68.0**	53.9	55.3
ODIN	51.8	50.5	50.1	50.0	55.3	56.3	64.5	51.5	51.4	52.2	68.0	53.8	54.6
Maha	49.3	39.1	34.5	**60.8**	50.0	50.1	49.2	51.6	50.1	49.9	48.2	50.0	48.6
kNN	49.9	50.0	50.0	50.0	49.8	50.6	51.7	45.4	48.4	50.4	45.8	50.2	49.4
ViM	45.2	36.8	33.6	50.0	48.6	49.8	56.3	50.3	51.6	46.3	42.6	49.9	46.7



(a) Best-detector counts **(b) Head-to-head coverage**

Figure 9: **Aggregate first place counts and pairwise coverage** over 48 dataset and backbone cells.

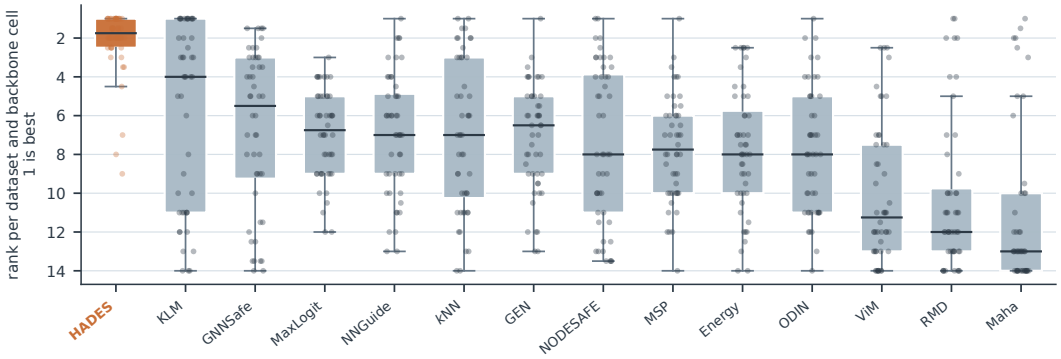


Figure 10: **Cellwise rank distributions** among HADES and 13 fixed detectors. Lower is better. Points show individual cells.

C.4 Full Per-Backbone Results

The following tables provide the dataset level evidence behind the aggregate summary. They report structural OOD AUROC, AUPR, and FPR95 for all 14 detectors on each of the four evaluated backbones. Tables 5 and 4 separately report the two model level references.

Table 12: **Structural OOD AUROC using the HGNN backbone** ($\times 100$). Entries are mean±standard deviation over three runs. Bold marks the best result in each dataset. The final columns report mean AUROC and average rank over 12 datasets.

Method	Cora-CA	Cora-CC	Cite.	PubM.	NTU	M.Net	Mush.	Sen.	House	DBLP	Zoo	20NG	Avg	Rank
HADES	**80.8±3.2**	**59.4±1.9**	**65.3±0.8**	81.1±1.4	**81.8±1.0**	**85.2±0.7**	**87.6±4.6**	43.1±6.3	58.2±3.1	**81.2±0.6**	61.3±14.3	65.7±1.2	70.9	2.08
GNNSafe	67.8±2.3	52.0±1.0	56.8±0.4	42.1±0.5	78.7±0.5	77.3±1.8	81.5±13.1	43.2±3.5	45.6±1.6	73.5±0.6	57.8±12.1	60.5±4.0	61.4	8.00
NODESAFE	68.4±3.3	51.9±0.9	56.1±0.5	42.1±0.5	71.4±1.4	69.6±0.7	70.4±5.2	44.5±2.0	47.5±1.4	69.4±0.5	57.8±12.1	55.6±3.8	58.7	9.00
Energy	50.6±10.6	51.8±1.4	56.0±0.5	47.7±7.5	75.5±0.4	77.4±1.0	71.3±16.9	47.7±3.8	47.4±2.8	70.3±6.8	42.0±3.7	57.8±3.0	58.0	8.75
NNGuide	76.5±3.1	59.4±1.0	59.0±0.7	42.4±0.3	80.2±0.4	80.7±0.2	74.9±12.3	47.4±4.7	51.9±4.2	79.5±0.4	42.7±11.6	63.5±2.1	63.2	5.08
MSP	72.2±1.3	54.9±0.7	57.3±0.4	42.2±0.5	76.4±0.8	77.7±0.2	83.4±2.0	39.0±4.9	54.0±2.8	79.6±0.2	56.6±8.8	**65.8±1.1**	63.3	5.62
ODIN	72.2±0.7	53.2±0.9	57.0±0.4	42.2±0.5	76.4±0.8	77.7±0.2	83.4±2.0	40.8±3.4	54.8±5.6	79.3±0.1	56.3±8.5	65.7±1.2	63.2	6.42
MaxLogit	72.4±1.3	55.1±0.7	57.3±0.4	42.2±0.5	78.8±0.6	79.5±0.2	77.7±8.7	41.4±9.6	52.2±2.1	79.9±0.3	56.7±8.9	63.7±1.8	63.1	5.33
GEN	71.7±1.5	54.1±0.7	57.1±0.4	42.2±0.5	78.7±0.7	78.5±0.1	83.4±2.0	38.9±4.8	54.0±2.8	80.2±0.2	**68.5±20.8**	65.3±1.3	64.4	5.46
KLM	23.0±3.3	23.1±0.1	18.4±1.3	**82.5±0.9**	55.3±0.2	67.2±9.1	69.5±8.5	46.3±5.8	49.4±5.8	57.7±0.5	59.4±25.8	45.9±3.6	49.8	9.58
kNN	**80.8±3.2**	59.2±1.7	60.6±0.8	39.4±0.6	76.1±1.4	85.1±0.0	59.9±7.3	35.8±8.0	49.1±9.3	**81.2±0.6**	40.8±19.1	48.9±4.9	59.7	7.67
Maha	21.1±1.2	26.9±1.1	19.7±1.1	81.1±0.1	42.9±1.3	48.7±0.7	58.1±2.3	**48.8±0.2**	48.4±1.9	40.1±1.0	48.3±4.4	41.3±5.3	43.8	10.83
RMD	25.8±0.5	27.8±0.9	20.2±0.5	4.5±4.2	43.0±0.4	48.8±0.2	61.5±2.9	41.0±17.2	**60.4±2.0**	43.2±0.2	51.6±1.6	53.0±1.2	40.1	10.58
ViM	23.4±4.4	29.2±1.6	25.5±1.3	47.7±7.5	35.2±0.8	41.0±1.2	71.3±16.9	47.7±3.8	47.4±2.8	20.0±0.6	48.3±1.6	47.0±2.6	40.3	10.58

Table 13: **Full structural OOD AUPR using the HGNN backbone.** Values are means over three runs ($\times 100$). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

AUPR ↑ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	**83.4**	**67.5**	**78.0**	88.3	76.7	84.3	**84.4**	48.8	55.6	**80.2**	65.7	57.3	72.5	2.67
GNNSafe	60.6	50.9	54.2	47.2	**79.0**	78.1	78.4	50.2	48.1	67.1	51.2	54.6	60.0	7.46
NODESAFE	61.7	50.5	53.8	47.2	72.4	71.4	71.0	52.0	50.1	60.2	51.2	53.1	57.9	9.21
Energy	54.6	51.7	54.0	49.8	76.1	77.3	71.5	52.9	46.8	69.5	53.9	53.5	59.3	8.00
NNGuide	79.4	65.3	58.7	47.7	78.5	78.9	76.2	**52.9**	51.6	74.1	54.2	55.5	64.4	4.50
MSP	67.8	54.1	54.8	47.2	74.8	76.3	79.8	46.4	52.1	74.1	59.2	57.3	62.0	6.33
ODIN	67.8	51.8	54.1	47.2	74.8	76.3	79.9	47.6	53.2	74.3	59.1	**57.4**	62.0	6.29
MaxLogit	68.2	54.3	54.7	47.2	77.3	78.0	79.6	49.1	50.8	74.2	60.9	55.7	62.5	5.38
GEN	67.2	53.2	54.5	47.2	77.2	77.3	79.8	46.3	52.1	74.4	**67.0**	56.8	62.8	5.83
KLM	41.9	44.5	43.0	**90.1**	58.6	69.7	62.9	51.4	49.0	62.5	63.0	45.4	56.8	9.33
kNN	**83.4**	66.1	66.4	44.4	76.9	**86.2**	55.3	46.4	48.1	**80.2**	52.8	44.7	62.6	7.33
Maha	40.6	47.1	44.6	89.8	44.2	54.5	58.9	48.9	47.9	47.8	51.7	43.3	51.6	11.17
RMD	41.4	47.1	45.2	44.9	47.2	50.6	59.3	48.2	**62.6**	46.2	58.8	55.2	50.6	10.33
ViM	36.4	45.2	47.2	49.8	43.8	48.0	71.5	52.9	46.8	36.3	50.7	47.3	48.0	11.17

Table 14: **Full structural OOD continued FPR95 using the HGNN backbone.** Values are means over three runs (×100). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

FPR95 ↓ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	100.0	98.6	**91.1**	98.5	**46.5**	55.1	**49.3**	95.3	87.9	55.2	96.2	60.8	77.9	4.29
GNNSafe	100.0	100.0	100.0	100.0	70.8	70.4	58.9	97.2	92.4	62.5	**44.9**	74.1	80.9	7.75
NODESAFE	100.0	100.0	100.0	100.0	79.8	82.5	89.6	97.2	93.0	65.8	**44.9**	78.0	85.9	9.17
Energy	**97.0**	**92.8**	96.2	**97.4**	73.0	69.6	82.1	98.1	96.5	78.1	97.4	94.0	89.4	7.54
NNGuide	100.0	100.0	100.0	100.0	55.9	**53.4**	78.7	97.7	90.5	53.8	91.0	73.1	82.8	7.04
MSP	100.0	100.0	100.0	100.0	61.4	61.8	63.1	94.8	87.9	50.6	84.6	60.5	80.4	6.04
ODIN	100.0	100.0	100.0	100.0	62.0	61.9	63.1	96.2	89.6	54.5	84.6	63.2	81.3	6.92
MaxLogit	100.0	100.0	100.0	100.0	59.4	59.2	82.7	98.1	94.0	49.9	89.7	67.6	83.4	7.79
GEN	100.0	100.0	100.0	100.0	61.3	61.3	63.1	95.8	88.2	**49.2**	62.8	**59.0**	78.4	5.75
KLM	97.6	94.8	93.0	97.7	97.3	94.0	92.4	97.2	93.4	95.2	80.8	74.8	92.4	7.50
kNN	100.0	100.0	100.0	100.0	77.9	60.4	83.1	94.4	**86.9**	55.2	92.3	74.0	85.4	7.42
Maha	99.0	93.6	91.5	99.4	95.2	97.0	94.2	**92.5**	92.3	98.4	100.0	93.4	95.5	7.88
RMD	100.0	99.8	100.0	97.9	100.0	100.0	97.4	94.8	98.8	100.0	100.0	100.0	99.1	10.88
ViM	97.6	97.9	93.7	**97.4**	99.9	99.8	82.1	98.1	96.5	99.6	98.7	98.5	96.7	9.04

Table 15: **Structural OOD AUROC using the UniGCNII backbone** ($\times 100$). Entries are mean±standard deviation over three runs. Bold marks the best result in each dataset. The final columns report mean AUROC and average rank over 12 datasets.

Method	Cora-CA	Cora-CC	Cite.	PubM.	NTU	M.Net	Mush.	Sen.	House	DBLP	Zoo	20NG	Avg	Rank
HADES	**72.0±1.4**	**57.5±1.0**	58.2±0.4	**80.6±0.0**	73.5±0.6	77.3±0.3	83.6±1.9	50.4±1.6	**51.7±0.1**	75.4±0.2	**73.8±3.9**	65.6±2.5	68.3	2.25
GNNSafe	**72.0±1.4**	54.8±1.1	57.6±0.4	42.1±0.5	71.6±1.2	72.1±0.0	83.6±1.9	44.2±2.2	47.8±0.9	75.4±0.2	64.6±7.3	65.6±2.5	62.6	7.04
NODESAFE	71.3±1.7	53.4±1.1	56.8±0.5	42.1±0.5	68.5±1.7	68.1±2.3	76.5±2.7	44.2±2.2	47.8±0.9	**76.9±0.3**	63.2±8.5	**65.7±3.2**	61.2	7.46
Energy	70.4±1.2	56.7±0.6	58.2±0.4	57.5±0.6	73.5±0.6	77.3±0.3	73.3±1.6	50.1±0.7	51.5±0.3	74.8±0.2	57.5±2.3	62.5±0.6	63.6	5.00
NNGuide	68.4±0.9	55.6±0.4	56.7±0.1	57.7±1.4	73.2±0.7	77.6±0.2	73.3±1.6	50.2±0.7	51.5±0.3	74.7±0.2	56.7±2.2	62.4±0.5	63.2	6.42
MSP	69.5±1.5	56.4±0.7	57.3±0.2	47.3±7.2	68.0±1.3	73.1±0.1	73.2±1.6	49.8±0.1	51.0±0.4	73.2±0.4	57.6±3.8	62.0±0.7	61.5	8.50
ODIN	71.1±1.7	57.4±1.0	**58.5±0.7**	47.8±7.2	68.2±1.2	73.3±0.0	73.1±1.6	49.7±0.1	51.0±0.4	70.9±1.6	57.6±3.7	62.2±0.7	61.8	7.08
MaxLogit	70.3±1.3	56.5±0.7	58.0±0.3	47.2±7.0	73.4±0.7	77.2±0.3	73.3±1.6	49.8±0.3	51.0±0.3	75.2±0.2	57.0±2.2	62.3±0.6	62.6	6.33
GEN	71.1±1.4	56.4±0.8	58.2±0.4	47.5±7.2	72.6±0.7	77.2±0.2	73.2±1.6	49.8±0.1	51.0±0.4	74.8±0.0	58.4±4.4	62.8±0.7	62.7	6.58
KLM	51.3±0.7	42.1±0.7	41.4±0.9	80.5±0.0	**79.3±0.8**	**90.3±0.2**	**87.1±0.7**	**50.8±0.7**	49.9±0.4	75.6±0.9	49.1±6.6	52.3±1.4	62.5	6.75
kNN	64.8±1.1	55.3±0.3	53.3±0.4	57.6±0.4	64.4±0.6	73.8±0.9	66.2±4.9	50.4±1.6	50.2±0.6	65.8±3.3	50.8±0.8	61.0±0.3	59.5	9.12
Maha	25.1±2.2	30.3±5.5	26.3±3.1	80.6±0.0	49.0±1.3	51.5±0.2	53.3±2.7	50.7±0.3	49.7±0.3	46.6±1.9	49.8±1.9	50.1±0.8	46.9	11.58
RMD	26.0±0.9	32.7±5.0	28.3±3.1	74.1±26.8	44.1±8.4	52.1±0.2	56.6±2.3	50.8±0.4	50.2±0.7	48.0±1.6	51.0±2.3	55.1±0.4	47.4	10.58
ViM	48.2±2.2	43.5±3.9	47.6±1.1	57.5±0.6	48.2±0.9	50.6±0.3	73.3±1.6	50.1±0.7	51.5±0.3	47.7±3.2	50.7±3.1	51.6±0.5	51.7	10.29

Table 16: **Full structural OOD AUPR using the UniGCNII backbone.** Values are means over three runs (×100). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

AUPR ↑ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	68.0	**58.1**	58.5	89.8	69.1	71.3	**83.3**	50.5	**51.5**	70.2	**69.7**	62.7	66.9	3.33
GNNSafe	68.0	54.9	57.8	47.2	70.0	70.8	**83.3**	**51.6**	50.4	70.2	68.2	62.7	62.9	5.42
NODESAFE	67.2	53.2	56.6	47.2	67.2	67.5	68.1	**51.6**	50.4	70.7	61.6	62.4	60.3	8.04
Energy	66.5	56.7	58.5	54.2	69.1	71.3	69.5	50.9	51.4	68.5	65.1	58.7	61.7	5.96
NNGuide	66.5	56.5	57.8	54.7	69.2	72.1	69.5	51.2	51.4	68.5	65.1	58.7	61.8	5.79
MSP	66.1	56.8	57.7	49.3	65.2	69.5	69.2	50.5	51.0	67.9	67.6	58.6	60.8	9.12
ODIN	68.0	58.1	**59.1**	50.5	65.5	69.9	69.1	50.4	51.0	68.0	67.6	58.8	61.3	6.83
MaxLogit	66.4	56.7	58.4	49.4	69.0	71.2	69.5	50.4	51.1	68.6	64.9	58.7	61.2	7.67
GEN	66.8	56.7	58.5	49.5	68.1	71.5	69.2	50.5	51.0	68.4	68.1	58.8	61.4	6.96
KLM	66.9	53.4	51.7	**90.2**	**85.3**	**93.6**	79.2	51.1	50.2	**80.2**	50.6	50.0	66.9	6.75
kNN	**70.3**	57.6	55.0	54.2	65.7	73.6	63.6	50.5	50.9	72.0	57.3	**64.2**	61.2	6.21
Maha	39.4	42.5	40.5	90.0	54.0	55.5	53.9	51.5	50.2	53.5	52.6	50.2	52.8	11.33
RMD	39.0	43.6	41.7	83.2	51.3	52.6	56.5	51.6	50.1	50.4	54.6	54.5	52.4	11.58
ViM	54.3	49.6	51.5	54.2	53.5	54.1	69.5	50.9	51.4	55.6	57.2	51.8	54.5	10.00

Table 17: **Full structural OOD continued FPR95 using the UniGCNII backbone.** Values are means over three runs ($\times 100$). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

FPR95 ↓ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	81.2	**93.5**	93.6	100.0	**67.4**	**57.2**	63.5	95.3	94.3	57.3	**59.0**	78.1	78.4	3.83
GNNSafe	81.2	94.1	94.0	100.0	72.5	71.2	63.5	97.2	**92.6**	57.3	82.0	78.1	82.0	5.50
NODESAFE	**80.0**	94.2	93.8	100.0	76.5	75.6	50.8	97.2	**92.6**	**54.6**	76.9	**77.6**	80.8	5.00
Energy	84.4	93.9	93.6	99.7	**67.4**	**57.2**	73.8	94.8	94.1	58.4	96.2	90.1	83.6	5.21
NNGuide	94.2	95.5	95.1	89.8	69.3	59.1	73.8	94.4	94.4	58.8	97.4	90.6	84.4	7.12
MSP	83.9	94.3	93.8	99.9	79.1	74.8	73.7	94.4	96.1	67.3	98.7	88.4	87.0	7.58
ODIN	87.8	93.7	**93.1**	97.4	79.5	74.4	73.4	94.4	96.0	84.2	98.7	88.3	88.4	6.50
MaxLogit	84.2	94.3	93.7	99.0	67.8	58.0	73.7	93.9	95.6	55.9	100.0	88.7	83.7	6.04
GEN	82.1	94.2	93.9	100.0	67.5	59.0	73.7	94.4	96.1	56.9	97.4	85.0	83.3	6.29
KLM	100.0	100.0	100.0	100.0	94.4	82.3	**29.8**	93.0	95.4	97.9	98.7	92.0	90.3	10.00
kNN	94.9	95.3	95.2	99.9	92.8	80.6	86.5	95.3	95.4	98.0	96.2	96.4	93.9	9.79
Maha	99.5	99.0	99.9	100.0	96.4	97.3	97.4	94.4	95.7	98.0	97.4	93.6	97.4	11.08
RMD	99.8	99.9	100.0	**55.6**	99.9	98.4	99.0	**92.5**	95.0	100.0	97.4	99.3	94.7	10.46
ViM	98.8	97.4	97.2	99.7	99.1	98.2	73.8	94.8	94.1	98.8	100.0	97.4	95.8	10.58

Table 18: **Structural OOD AUROC using the AllSetTr. backbone** ($\times 100$). Entries are mean±standard deviation over three runs. Bold marks the best result in each dataset. The final columns report mean AUROC and average rank over 12 datasets.

Method	Cora-CA	Cora-CC	Cite.	PubM.	NTU	M.Net	Mush.	Sen.	House	DBLP	Zoo	20NG	Avg	Rank
HADES	**68.6±1.9**	53.3±0.4	**56.2±0.7**	**81.2±0.1**	63.5±0.4	68.6±0.8	76.2±3.0	**52.1±1.4**	**53.2±2.0**	70.7±1.6	**78.0±6.3**	**67.0±1.3**	65.7	1.62
GNNSafe	**68.6±1.9**	52.1±0.7	**56.2±0.7**	42.1±0.5	61.5±1.3	57.3±0.1	72.7±3.1	44.0±2.3	47.7±1.5	69.2±0.4	50.1±11.7	**67.0±1.3**	57.4	5.54
NODESAFE	67.8±2.4	51.9±0.8	55.8±0.4	42.1±0.5	61.2±1.7	57.1±0.2	61.5±2.8	44.2±2.0	47.7±1.5	69.0±0.4	54.8±8.8	59.0±2.8	56.0	6.50
Energy	52.9±0.2	51.8±1.0	50.7±0.6	46.7±0.6	57.7±0.4	53.7±0.1	58.2±1.1	49.1±1.5	48.6±0.9	52.4±0.3	45.0±5.2	59.4±2.0	52.2	8.21
NNGuide	53.2±0.2	51.9±1.0	50.7±0.6	45.6±1.0	58.7±0.3	54.6±0.1	58.2±1.1	49.5±1.2	48.6±0.9	52.1±0.2	45.3±5.1	59.4±2.0	52.3	8.04
MSP	52.9±0.5	51.6±1.0	50.9±0.4	51.6±3.0	56.0±0.3	52.3±0.3	59.0±1.8	50.6±1.1	48.1±2.1	51.5±0.4	46.8±7.6	58.6±0.8	52.5	8.50
ODIN	51.6±0.8	51.0±0.7	49.4±0.8	50.3±2.6	55.8±0.3	52.2±0.3	59.0±1.8	50.8±0.9	48.2±2.1	49.6±1.1	46.4±8.3	58.1±0.7	51.9	9.58
MaxLogit	53.3±0.3	51.7±1.0	50.9±0.5	49.8±3.0	57.6±0.2	53.7±0.1	58.2±1.1	50.2±1.6	47.9±1.8	52.4±0.3	46.6±6.4	59.3±1.9	52.6	8.08
GEN	53.3±0.2	51.7±0.9	51.0±0.4	50.0±0.0	57.1±0.4	52.9±0.2	59.0±1.8	50.6±1.1	48.1±2.1	52.0±0.4	44.9±5.4	58.8±0.8	52.5	7.92
KLM	56.5±1.1	**54.2±1.4**	50.3±1.1	46.3±1.6	**70.3±1.0**	67.6±0.9	**89.2±11.0**	52.0±6.5	51.1±12.0	**85.6±8.7**	52.7±4.2	60.1±0.6	61.3	3.83
kNN	56.1±0.3	53.3±0.4	49.5±2.6	47.8±0.9	63.5±0.4	**68.7±1.1**	59.1±2.0	52.0±3.7	52.9±2.3	57.4±0.9	48.5±5.4	60.3±1.2	55.8	4.46
Maha	41.9±3.4	44.0±5.3	45.6±4.4	44.1±27.8	54.7±0.4	55.4±0.3	56.1±2.0	**52.1±1.4**	46.0±8.5	32.4±0.4	44.8±1.8	58.8±1.5	48.0	11.17
RMD	41.8±3.5	44.1±5.6	45.5±3.9	79.7±3.3	52.9±0.6	52.4±0.4	48.8±1.6	50.5±1.2	42.6±17.5	30.6±1.1	47.4±1.0	55.5±0.9	49.3	11.42
ViM	50.0±0.4	49.9±0.5	51.0±0.6	46.7±0.6	50.9±0.5	51.0±0.4	58.2±1.1	49.1±1.5	48.6±0.9	49.3±0.8	54.6±6.9	51.8±0.4	50.9	10.12

Table 19: **Full structural OOD AUPR using the AllSetTr. backbone.** Values are means over three runs ($\times 100$). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

AUPR ↑ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	**62.9**	53.5	**55.9**	**89.3**	65.7	70.1	69.4	**54.4**	**53.1**	65.1	**76.0**	**64.8**	65.0	1.71
GNNSafe	**62.9**	51.9	**55.9**	47.2	58.5	55.4	59.2	50.7	50.3	61.3	56.9	**64.8**	56.2	4.96
NODESAFE	61.1	51.5	55.6	47.2	58.2	55.2	52.9	51.7	50.3	60.8	56.8	57.1	54.9	6.42
Energy	51.6	51.4	50.6	48.7	55.6	52.6	56.3	50.9	49.6	51.5	51.5	56.1	52.2	8.71
NNGuide	51.9	51.6	50.6	48.1	56.7	53.3	56.3	50.9	49.7	51.4	51.5	56.1	52.3	8.38
MSP	52.2	51.2	50.6	51.2	55.0	51.9	55.9	51.8	48.1	50.9	52.7	55.3	52.2	8.92
ODIN	51.9	50.8	49.7	50.9	55.0	51.9	55.9	52.7	48.1	49.9	52.1	55.1	52.0	9.88
MaxLogit	52.0	51.3	50.5	50.3	55.6	52.5	56.3	52.5	49.3	51.5	52.0	56.0	52.5	7.88
GEN	52.2	51.0	50.7	50.0	55.7	52.1	55.8	52.0	48.1	50.3	52.1	55.4	52.1	8.88
KLM	57.8	54.0	50.3	48.7	**73.8**	**72.7**	**89.2**	54.2	49.9	**84.8**	57.0	55.5	62.3	4.08
kNN	57.3	**54.6**	49.8	49.2	65.7	70.1	56.0	52.2	52.7	57.8	53.5	60.7	56.6	5.12
Maha	45.1	46.8	47.4	65.2	60.1	59.2	54.9	**54.4**	47.6	42.0	47.1	57.4	52.3	9.21
RMD	45.0	47.0	47.2	80.6	54.9	53.0	47.2	51.5	47.7	43.8	56.2	55.4	52.5	10.92
ViM	51.0	50.0	51.2	48.7	53.6	51.4	56.3	50.9	49.6	47.5	60.1	54.1	52.0	9.96

Table 20: **Full structural OOD continued FPR95 using the AllSetTr. backbone.** Values are means over three runs ($\times 100$). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

FPR95 ↓ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	80.2	94.9	93.9	99.4	92.0	90.4	60.8	93.9	95.5	70.0	**64.1**	**78.0**	84.4	5.75
GNNSafe	80.2	94.6	93.9	100.0	81.2	83.2	43.9	96.2	93.0	**64.0**	89.7	**78.0**	83.2	4.67
NODESAFE	**79.8**	94.2	**93.9**	100.0	**81.0**	**83.0**	53.6	96.7	93.0	64.3	85.9	**78.0**	83.6	3.75
Energy	93.4	94.5	94.8	97.9	86.8	92.3	86.0	94.4	94.5	92.2	100.0	86.6	92.8	7.25
NNGuide	93.6	94.6	94.9	98.4	87.1	91.8	86.0	93.9	94.5	93.7	96.2	88.2	92.7	7.46
MSP	93.8	94.5	94.9	93.0	90.8	93.9	86.0	93.4	94.5	93.6	98.7	85.6	92.7	6.92
ODIN	94.6	94.3	95.4	94.2	91.8	94.2	86.1	93.4	93.6	95.1	98.7	86.2	93.1	7.88
MaxLogit	93.4	**94.1**	94.8	95.7	87.6	92.5	86.0	93.9	95.0	92.3	100.0	84.9	92.5	6.58
GEN	93.3	94.4	95.1	100.0	86.3	92.6	86.0	93.4	94.5	93.3	100.0	85.8	92.9	6.96
KLM	94.0	94.5	95.5	98.7	89.6	92.2	**18.3**	95.3	**81.8**	85.9	96.2	94.9	86.4	7.33
kNN	93.8	94.3	95.2	97.0	92.0	91.2	89.7	**92.0**	95.2	95.5	97.4	95.3	94.1	8.17
Maha	100.0	98.2	98.1	94.9	99.1	96.3	93.0	93.9	95.8	100.0	94.9	96.3	96.7	11.21
RMD	100.0	98.4	98.1	**71.0**	99.3	96.4	92.6	94.4	96.3	100.0	94.9	98.7	95.0	11.58
ViM	95.3	95.1	95.4	97.9	96.1	96.0	86.0	94.4	94.5	88.8	93.6	99.7	94.4	9.50

Table 21: **Structural OOD AUROC using the ED-HNN backbone** ($\times 100$). Entries are mean±standard deviation over three runs. Bold marks the best result in each dataset. The final columns report mean AUROC and average rank over 12 datasets.

Method	Cora-CA	Cora-CC	Cite.	PubM.	NTU	M.Net	Mush.	Sen.	House	DBLP	Zoo	20NG	Avg	Rank
HADES	**80.6±7.3**	53.2±1.3	**57.1±0.3**	**81.5±1.0**	58.3±1.3	67.1±1.6	**88.9±2.6**	47.2±4.2	54.8±1.8	76.2±1.3	**65.4±11.3**	68.8±4.1	66.6	2.67
GNNSafe	67.8±2.1	52.0±0.8	56.1±0.4	42.1±0.5	57.5±1.4	56.3±0.1	64.1±9.7	44.2±2.2	48.5±1.2	68.8±0.2	58.4±14.7	70.2±3.1	57.2	6.29
NODESAFE	67.8±2.1	51.9±0.9	56.1±0.3	42.1±0.5	57.5±1.4	56.3±0.0	68.6±6.7	44.2±2.1	47.6±1.1	68.8±0.2	58.4±14.7	57.9±3.9	56.4	6.88
Energy	39.1±4.8	50.0±3.7	53.4±3.2	52.1±6.6	46.9±1.6	55.8±4.2	51.4±9.2	52.6±2.4	50.3±1.8	32.8±0.6	49.8±4.3	69.7±4.0	50.3	9.25
NNGuide	39.7±8.0	49.0±4.3	54.3±1.9	51.8±6.8	58.2±4.5	57.5±4.3	51.4±9.2	**52.7±2.4**	50.3±1.9	32.9±0.6	51.3±4.2	69.7±4.0	51.6	8.21
MSP	47.7±16.1	47.9±1.8	49.3±5.7	42.1±0.7	47.7±0.9	60.2±3.0	57.6±9.8	50.0±0.7	51.0±0.1	44.7±10.7	54.2±2.9	66.9±0.7	51.6	8.08
ODIN	42.1±8.8	46.8±2.2	49.0±5.9	47.2±6.9	47.6±1.1	60.1±3.1	57.3±9.7	50.1±0.8	51.0±0.3	40.9±11.2	54.5±3.3	67.0±0.7	51.1	8.42
MaxLogit	47.2±15.9	48.7±3.8	51.1±4.7	52.2±6.8	47.2±0.7	59.2±3.7	51.5±9.1	52.4±1.8	50.6±1.5	33.4±0.8	53.1±3.6	69.5±3.3	51.3	8.00
GEN	46.9±16.3	51.1±0.6	48.6±6.1	52.7±7.1	46.1±0.5	52.0±5.3	57.6±9.8	50.0±0.7	51.0±0.1	44.1±14.5	51.3±5.9	68.1±0.9	51.6	8.58
KLM	80.2±6.2	**56.0±4.9**	50.5±14.2	80.5±0.0	**76.1±3.3**	**90.0±6.1**	67.0±25.0	50.4±2.7	52.3±2.3	**99.7±0.4**	53.2±3.7	77.5±5.5	69.4	3.12
kNN	58.2±13.5	50.9±1.9	53.0±3.8	43.7±7.9	58.8±1.3	67.0±1.4	56.7±4.5	46.5±3.3	**54.9±1.7**	65.3±2.7	52.7±2.0	64.2±0.8	56.0	6.83
Maha	63.4±13.3	51.8±1.6	47.8±6.1	80.5±0.0	27.4±4.6	40.1±7.5	45.4±4.9	44.1±7.0	46.0±1.5	42.4±7.1	47.3±3.0	65.5±1.1	50.2	10.88
RMD	40.7±8.5	47.2±6.0	36.6±9.7	62.0±42.8	47.1±5.6	55.8±6.4	72.8±7.1	46.4±2.6	50.6±0.7	68.5±3.1	50.8±2.7	**79.4±2.9**	54.8	8.42
ViM	53.4±6.0	49.7±1.9	54.0±2.6	52.1±6.6	45.5±5.2	42.1±1.3	51.4±9.2	52.6±2.4	50.3±1.8	45.7±2.9	47.5±1.3	54.5±4.0	49.9	9.38

Table 22: **Full structural OOD AUPR using the ED-HNN backbone.** Values are means over three runs (×100). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

AUPR ↑ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	82.4	55.0	**56.1**	**89.9**	57.5	70.0	**85.6**	49.5	55.1	81.6	**60.9**	69.3	67.8	3.17
GNNSafe	61.4	51.4	55.5	47.2	54.0	54.0	56.8	51.9	49.7	59.3	51.8	67.9	55.1	7.62
NODESAFE	61.4	51.4	55.5	47.2	54.0	54.1	57.3	51.7	50.2	59.2	51.8	56.2	54.2	7.88
Energy	41.5	50.0	51.7	52.8	46.0	54.9	50.9	**53.2**	50.6	38.4	51.6	71.8	51.1	9.29
NNGuide	42.0	49.0	53.2	52.3	58.0	56.6	50.9	**53.2**	50.6	38.4	52.6	71.9	52.4	7.83
MSP	48.2	48.4	49.8	46.8	46.6	56.9	56.6	51.4	50.6	44.5	57.1	67.2	52.0	9.04
ODIN	44.2	47.7	49.6	49.9	46.5	56.8	56.4	51.5	50.6	43.0	57.4	67.4	51.7	9.00
MaxLogit	47.8	49.2	50.5	52.7	46.2	56.4	51.0	53.1	51.3	38.6	55.5	71.9	52.0	7.92
GEN	47.5	50.3	49.5	51.9	46.0	51.0	56.7	51.4	50.6	45.1	53.6	67.5	51.7	9.54
KLM	**85.7**	**60.3**	51.9	89.8	**76.7**	**91.5**	71.2	53.0	52.2	**99.8**	56.5	74.4	71.9	2.54
kNN	58.3	56.2	53.1	49.3	59.0	69.8	54.1	51.1	**55.6**	60.0	55.3	63.6	57.1	6.00
Maha	69.4	54.0	47.9	89.8	36.6	46.3	45.7	48.1	48.7	46.5	51.0	69.7	54.5	9.88
RMD	45.9	51.6	44.7	77.4	55.2	66.2	69.5	48.1	51.4	70.3	54.1	**80.0**	59.5	6.17
ViM	50.5	49.5	52.9	52.8	53.4	43.3	50.9	**53.2**	50.6	44.1	51.4	53.6	50.5	9.12

Table 23: **Full structural OOD continued FPR95 using the ED-HNN backbone.** Values are means over three runs ($\times 100$). Bold marks the best result in each column. The final columns report the mean and average rank over 12 datasets.

FPR95 ↓ **Method**	**Cora-CA**	**Cora-CC**	**Cite.**	**PubM.**	**NTU**	**M.Net**	**Mush.**	**Sen.**	**House**	**DBLP**	**Zoo**	**20NG**	**Avg**	**Rank**
HADES	91.8	**94.1**	95.1	99.2	80.7	83.4	**29.6**	94.8	**91.6**	81.2	74.4	82.5	83.2	4.42
GNNSafe	**80.7**	94.2	94.0	100.0	83.2	83.5	54.4	96.2	92.5	66.1	**53.9**	66.7	80.4	4.21
NODESAFE	80.8	94.2	94.0	100.0	83.2	83.4	43.3	96.2	93.1	66.1	**53.9**	78.0	80.5	4.50
Energy	96.2	94.8	94.0	96.3	95.0	89.0	87.5	94.4	94.6	97.1	93.6	88.9	93.4	7.92
NNGuide	96.6	95.2	95.2	96.5	84.6	87.0	87.5	93.9	95.2	97.0	94.9	89.1	92.7	8.92
MSP	91.2	95.0	94.7	100.0	96.6	85.2	85.3	97.7	94.4	88.2	93.6	89.5	92.6	8.75
ODIN	96.8	94.7	94.7	97.8	96.5	84.9	85.3	97.7	94.1	95.3	93.6	89.5	93.4	8.21
MaxLogit	89.5	95.1	93.7	94.5	96.1	84.7	87.3	94.8	95.0	97.0	91.0	92.8	92.6	7.46
GEN	89.6	94.3	94.2	99.6	96.8	88.7	85.3	97.7	94.4	86.9	98.7	81.5	92.3	8.04
KLM	99.6	95.7	**87.3**	100.0	**69.8**	**42.9**	47.7	**93.0**	94.7	**0.2**	97.4	**51.9**	73.3	5.83
kNN	97.9	95.3	95.3	99.7	79.0	82.9	87.7	93.9	**91.6**	80.1	94.9	86.2	90.4	6.96
Maha	97.7	94.7	96.9	100.0	84.5	90.0	93.7	93.4	94.3	91.5	98.7	88.4	93.7	9.46
RMD	98.1	97.2	98.7	**70.4**	96.8	97.7	91.3	95.8	93.3	99.8	98.7	96.0	94.5	11.33
ViM	91.2	94.9	96.4	96.3	88.0	91.0	87.5	94.4	94.6	90.5	97.4	100.0	93.5	9.00

D Extended Related Work

Post-hoc detection. Confidence and logit scores are sensitive to changes in decision margin (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Hendrycks et al., 2022; Liu et al., 2023), whereas feature-distance scores detect changes relative to training geometry (Lee et al., 2018; Ren et al., 2021; Sun et al., 2022; Wang et al., 2022; Park et al., 2023). Structural perturbations can manifest through either path after message passing, which explains why the candidate ordering depends on the backbone as well as the dataset.

Relational confidence. GNNSafe (Wu et al., 2023) improves node OOD detection by propagating confidence over an in-distribution graph. NODESAFE (Yang et al., 2024) regularizes energy during training. HyperGOOD (Cai et al., 2026) brings multiscale representations and energy propagation to hypergraphs but restricts its reported benchmark to feature and label perturbations under fixed incidence. HADES asks the complementary question: how should a detector behave when the incidence used for propagation is itself shifted?

Learned incidence. DHGNN (Jiang et al., 2019) and HERALD (Zhang et al., 2022) learn or refine hypergraph structure to improve classification. HADES does not modify the trained backbone. It builds a post-hoc prototype incidence whose centers are anchored to clean train embeddings and uses the incidence only to construct an alternative propagation view.