Simplex Regularization Improves Out-of-Distribution Detection on Graphs

A.Q.M. Sazzad Sayyed[†], Nathaniel D. Bastian[‡] and Francesco Restuccia[†]

[†]Northeastern University [‡]United States Military Academy

Abstract

Node-level graph Out-of-Distribution (OOD) detection is challenging because message passing smooths node scores across neighborhoods, making In-Distribution (ID) and OOD samples appear similar. Existing propagation-based detectors improve detection by regularizing energy scores but overlook classifier geometry. As such, we introduce Collapse Safe (CSafe), a simplex-Equiangular Tight Frame (ETF) regularizer that aligns the logit space with uniformly separated prototypes of learnable norm while retaining the standard energy-propagation pipeline. This isolates the effect of classifier geometry from changes to the detector. The objective concentrates class outputs around separated prototypes, reduces class-dependent energy variation, and improves propagated ID/OOD separability. Across 20 dataset-shift pairs on eight node-level graph OOD benchmarks, CSafe outperforms the state-of-the-art NodeSafe on 17 pairs. On Cora, Citeseer, and Pubmed, it reduces False Positive Rate (FPR) by 4.82% with a Graph Convolution Network (GCN), showing that representation geometry is a complementary lever to information propagation for graph OOD detection.

1 Introduction

Graph neural networks (GNNs) are employed in critical domains such as citation analysis (Kipf and Welling 2017), recommendation systems (Ying et al. 2018), cyber-security (Bilot et al. 2023) and biological networks (Zitnik, Agrawal, and Leskovec 2018). Since GNNs usually operate under open-world assumption, they must differentiate ID data from OOD data – a problem known as graph out-of-distribution (G-OOD) detection (Liang, Li, and Srikant 2017). This is significantly harder than conventional OOD detection, since graph nodes are coupled through message passing, thus violating the independence assumption of traditional methods (Zhao et al. 2020; Ma, Deng, and Mei 2021; Wu et al. 2022).

Key Issues. Recent G-OOD methods have adopted information propagation(Wu et al. 2023a; Yang et al. 2024; Bao et al. 2024) to improve ID/OOD separation. GNNSafe (Wu et al. 2023a) propagates node-level energy scores to better distinguish ID from OOD, while NodeSafe (Yang et al. 2024) further enhances stability by constraining the logits, which bounds extreme energy values during propagation. *Both methods, however, overlook the geometry of the representation space*. When that space is poorly structured, propagation smooths incompatible energy values across neighboring samples, degrading ID/OOD separability regardless of any score-level regularization.

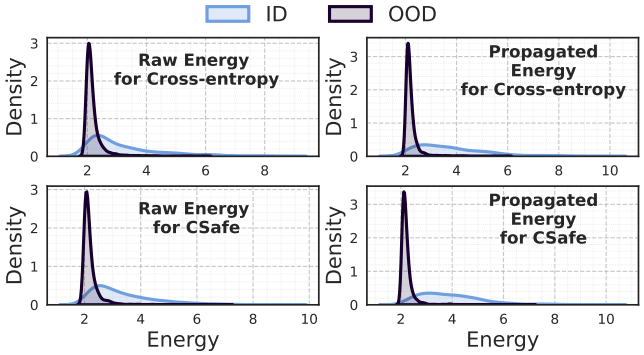


Figure 1: The top two figures show the score distribution before and after propagation for Cora-label OOD. The bottom two figures show the same for CSafe. CSafe shifts the ID energy even before the energy propagation – creating a logit-space more suitable for ID/OOD separation.

Proposed Approach. Motivated by recent evidence that neural-collapse geometry can improve OOD detection in image classification (Ammar et al. 2024; Harun, Gallardo, and Kanan 2025; Sayyed, Bastian, and Restuccia 2026), we ask: *Can inducing a simplex-structured classifier output space improve the separation of propagated energy scores?* To answer this question, we propose Collapsed Softmax (CS) as a training-time regularizer that encourages ID node logits to align with fixed simplex ETF prototypes. Inference is unchanged from GNNSafe and NodeSafe: we use the same energy scores and the same propagation mechanism. *This controlled design isolates the contribution of representation geometry from that of the detector itself*. Our results answer the research question affirmatively: even partial alignment to simplex-ETF yields measurable improvements in OOD detection. We prove that CSafe induces a margin-controlled concentration bound on ID energy scores, addressing a failure mode of energy propagation identified by (Yang et al. 2024). Under K steps of scalar energy propagation, the bound decomposes into a representation-

controlled ID concentration term and a topology-controlled ID/OOD score-contamination term. This analysis characterizes when CSafe's geometric regularization is preserved and when its benefits are limited by graph topology. Across 20 dataset/OOD-type pairs spanning seven benchmark graphs and feature, structural, and label shifts, CSafe outperforms NodeSafe in 16 of 20 cases by Area Under Receiver Operating Curve (AUROC). On the more challenging Cora, Citeseer, and Pubmed benchmarks, it improves AUROC over SOTA Nodesafe by 1% on average and reduces FPR by 4.82% on average, with maximum gains exceeding 10%.

Summary of Key Contributions

- We identify partial simplex alignment of the logits (and, consequently, the learned features) as an underexplored mechanism for graph OOD detection that complements energy-based regularization. To exploit this mechanism, we introduce CS regularization, which improves detection without altering inference and adds negligible number of trainable parameters with little increase in computational overhead.
- We derive a rigorous margin-dependent norm and an exact common energy for all canonical ETF prototypes. We further show that, after K scalar propagation steps, the ID energy deviation decomposes into representation-controlled concentration and topology-controlled ID/OOD score contamination.
- Through extensive experiments across 8 benchmarks and 3 OOD-types (structural, feature, label), we show that CSafe improves over NodeSafe on 16 of 20 dataset and OOD-type pairs, with the strongest gains on co-purchase and co-authorship families. Specifically, for Cora, Citeseer, and Pubmed benchmarks, CSafe improves over the closest competitor NodeSafe by 4.82% in terms of FPR on average.

The remainder of the paper is organized as follows. Section 2 reviews related works on graph OOD detection and neural collapse. Section 3 formalizes the node-level graph OOD setup and the energy-propagation pipeline used by prior methods. Section 4 introduces CSafe along with its components and end-to-end methodology. Section 5 gives a theoretical analysis. Section 6 reports experiments across benchmarks and ablations. Section 7 concludes with limitation and future work.

2 Related Work

Graph OOD Detection. OOD detection has been widely studied in i.i.d. domains using both logit and feature based scoring methods such as MSP (Hendrycks and Gimpel 2016), ODIN (Liang, Li, and Srikant 2017), Mahalanobis distance (Lee et al. 2018) and energy-based detection (Liu et al. 2020). G-OOD detection is substantially more challenging, since node representations are coupled through message passing and neighborhood aggregation, leading to oversmoothing of the energy score and reduced separation of the ID and OOD nodes. Most of existing G-OOD approaches focusing on graph classification settings (Li et al. 2022; Bazhenov et al. 2022; Liu et al. 2023a; Guo et al. 2023; Ding and Shi 2023). SIGNET (Liu et al. 2023b) uses unsupervised learning for anomaly detection. For node-level G-OOD, GKDE (Zhao et al. 2020) estimates graph-aware uncertainty using Dirichlet distributions, while Graph Posterior Networks (Stadler et al. 2021) perform Bayesian posterior estimation over graph representations. More recently, GNNSAFE (Wu et al. 2023a) introduced graph energy propagation for node-level OOD detection, and NODESAFE (Yang et al. 2024) further stabilized propagated energies through boundedness and variance regularization. (Bao et al. 2024) further incorporates topology information in the OOD score. *Our work separates itself since it focuses on how representation geometry itself influences propagated energy behavior.*

Neural Collapse, Simplex Geometry in G-OOD. Neural collapse describes the phenomenon where deep classifiers converge toward simplex ETF structures during late-stage training (Papyan, Han, and Donoho 2020). Recent works exploit simplex-based classifiers and ETF-inspired objectives to improve optimization, robustness and calibration (Li et al. 2023). While neural collapse has been extensively studied in i.i.d. settings, its interaction with graph message passing and G-OOD detection remains largely unexplored. Unlike conventional settings, graph message passing continuously smooths node representations across neighborhoods, potentially perturbing the induced simplex geometry. Our work connects these directions by showing that *approximate simplex collapse can improve graph OOD detection by inducing more stable propagated energy landscapes*.

3 Background and Problem Setup

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ denote an attributed graph, where $\mathcal{V}$ is the node set, $\mathcal{E}$ is the edge set, and $\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$ contains node features. A GNN encoder f_θ computes node representations through iterative message passing:

$$h_i^{(l+1)} = \phi\Big(h_i^{(l)}, \mathrm{AGG}\Big(\Big\{h_j^{(l)} \mid j \in \mathcal{N}(i)\Big\}\Big)\Big), \quad (1)$$

where $\mathcal{N}(i)$ is the neighborhood of node v_i, $\mathrm{AGG}(\cdot)$ is a neighborhood aggregation operator, and $\phi(\cdot)$ is an update function. We denote the final representation after L layers as $h_i = h_i^{(L)}$. A linear classifier maps h_i to logits

$$z_i = W h_i + b, \quad (2)$$

where $W \in \mathbb{R}^{C \times d}$, $b \in \mathbb{R}^C$, and C is the number of ID classes.

OOD detection as node-level hypothesis testing. The goal is to decide whether a test node v_i belongs to the ID distribution $\mathcal{P}_{id}^{\mathcal{V}}$ or the OOD distribution $\mathcal{P}_{ood}^{\mathcal{V}}$. Because a GNN computes h_i using both node attributes and neighborhood context, the OOD score is graph-dependent. We define a scalar scoring function $S : \mathcal{V} \times \mathcal{G} \to \mathbb{R}$ inducing

$$\begin{aligned} S(v_i, \mathcal{G}) &\sim \mathcal{P}_{id}^{S} \quad \text{if } v_i \sim \mathcal{P}_{id}^{\mathcal{V}}, \\ S(v_i, \mathcal{G}) &\sim \mathcal{P}_{ood}^{S} \quad \text{if } v_i \sim \mathcal{P}_{ood}^{\mathcal{V}}. \end{aligned} \quad (3)$$

A node-level decision is obtained by thresholding:

$$\mathbb{1}(v_i, \mathcal{G}) = \begin{cases} 1, & \text{if } S(v_i, \mathcal{G}) \geq \tau, \\ 0, & \text{if } S(v_i, \mathcal{G}) < \tau, \end{cases} \quad (4)$$

where $\mathbb{1} = 1$ indicates ID and $\mathbb{1} = 0$ indicates OOD. The threshold τ controls the True Positive Rate (TPR)/FPR trade-off. Performance is evaluated by fixing a target TPR over ID test nodes and reporting the corresponding FPR over OOD test nodes. We assume throughout that higher scores indicate stronger ID confidence.

4 The CSafe Methodology

Structuring the Energy Landscape

Node-level graph OOD detection often relies on a scalar confidence score computed from the GNN output. Let $u_i = f_\theta(x_i, \mathcal{G}) \in \mathbb{R}^C$ denote the output logits for node i. We use the negative energy score (Liu et al. 2020) $S_i = T \log \sum_{c=1}^{C} \exp(u_{i,c}/T)$ as the OOD score, where $T > 0$ is the temperature. Larger S_i indicates stronger ID confidence.

Graph-based detectors further propagate this score over the graph:

$$\widetilde{S}^{(k+1)} = \eta \widetilde{S}^{(k)} + (1-\eta) P \widetilde{S}^{(k)}, \qquad \widetilde{S}^{(0)} = S, \tag{5}$$

where P is a normalized propagation matrix and $\eta \in [0, 1]$ controls the amount of self-retention. For one propagation step,

$$\widetilde{S}_i - S_i = (1-\eta) \left(\sum_j P_{ij} S_j - S_i \right). \tag{6}$$

Thus, each node score is pulled toward its neighborhood average. This can improve local consistency, but it can also oversmooth the energy landscape as illustrated in (Yang et al. 2024). An OOD node connected to confident ID nodes may inherit an ID-like score. Similarly, an ID node near ambiguous or topology-mixed regions may receive an unstable score. Existing graph energy methods mainly operate after the scalar energy has already been computed. In contrast, CSafe regularizes the logit space that generates the energy. The goal is to make class logits geometrically structured before energy propagation is applied. This encourages compact class-aligned outputs, more uniform class separation, and a smoother pre-propagation energy landscape.

Simplex-ETF Prototypes

CSafe uses a simplex-ETF as a target geometry for the GNN outputs. Let C be the number of classes. We define the matrix of canonical class prototypes $q_1, \ldots, q_C \in \mathbb{R}^C$ by

$$Q = \sqrt{\frac{C}{C-1}} \left(I_C - \frac{1}{C} \mathbf{1}\mathbf{1}^\top \right), \tag{7}$$

where the c-th row of Q gives q_c. These prototypes satisfy

$$q_c^\top q_k = \begin{cases} 1, & c = k, \\ -\dfrac{1}{C-1}, & c \neq k. \end{cases} \tag{8}$$

Thus, all class prototypes are equally separated. This removes arbitrary class-wise geometric imbalance from the target logit space. Rather than fixing the prototype radius, CSafe learns a positive scale parameter. Let r be an unconstrained scalar parameter. The effective prototype norm is

$$\rho = \mathrm{softplus}(r), \tag{9}$$

and the scaled prototype for class c is

$$p_c = \rho \, q_c. \tag{10}$$

The softplus parameterization keeps the radius positive. In implementation, the gradient of ρ is scaled by a small constant to prevent instability in prototype radius updates. Apart from that, in the two-stage training pipeline (as described in Subsection *Training Objective*), r is initialized with the mean of the logit norm after the end of the first stage.

Distance-Based CS Logits

For a labeled training node (i, y_i), CS computes squared distances between the GNN output u_i and each prototype:

$$d_{i,c} = \|u_i - p_c\|_2^2 . \tag{11}$$

To strengthen the assigned-class constraint, CS applies a multiplicative margin to the target distance:

$$\widetilde{d}_{i,c} = \begin{cases} \exp(m)\, d_{i,y_i}, & c = y_i, \\ d_{i,c}, & c \neq y_i, \end{cases} \tag{12}$$

where $m \geq 0$ is the margin parameter. The corresponding CS logits are

$$z_{i,c}^{\mathrm{CS}} = -s \, \widetilde{d}_{i,c}, \tag{13}$$

where $s > 0$ is a distance-logit scale.

The CSafe regularization loss is the cross-entropy over these distance logits:

$$\mathcal{L}_{\mathrm{CS}} = -\log \frac{\exp(-s \exp(m) d_{i,y_i})}{\exp(-s \exp(m) d_{i,y_i}) + \sum_{c \neq y_i} \exp(-s d_{i,c})}. \tag{14}$$

This objective encourages u_i to lie closer to its assigned simplex prototype than to other prototypes. Since the prototypes are equiangular, this also encourages balanced inter-class separation.

Importantly, the CSafe logits are used only for training. They are not used as the test-time detector. At inference, the model uses the original GNN logits u_i and the standard negative-energy score.

Training Objective

The base supervised loss is the standard cross-entropy over the GNN logits:

$$\mathcal{L}_{\mathrm{CE}} = -\log \frac{\exp(u_{i,y_i})}{\sum_{c=1}^{C} \exp(u_{i,c})}$$

CS combines this classification loss with the simplex-distance regularizer into a single weighted objective,

$$\mathcal{L} = \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CS}} \mathcal{L}_{\mathrm{CS}}, \tag{15}$$

where λ_{CE} preserves ordinary logit semantics and λ_{CS} controls the strength of the simplex geometry regularization. Rather than optimizing this combined objective from a random initialization, we realize it through a two-stage training procedure so that the simplex-ETF geometry is imposed on an already class-discriminative logit space rather than shaping it from scratch.

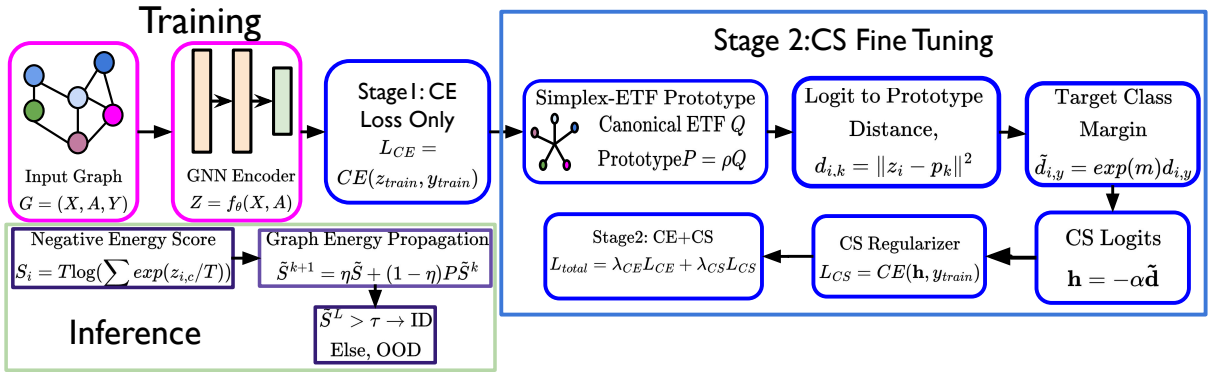


Figure 2: An overview of the two-stage training and inference pipeline of CSafe.

Stage 1: Cross-entropy pretraining. We first train the GNN encoder using the standard supervised objective

$$\mathcal{L}^{(1)} = \mathcal{L}_{\text{CE}} = \text{CE}(Z_{\mathcal{I}_{\text{tr}}}, Y_{\mathcal{I}_{\text{tr}}}), \qquad Z = f_\theta(X, A). \quad (16)$$

This stage learns a conventional classifier whose logits remain compatible with energy-based detection.

Stage 2: CS finetuning. Starting from the cross-entropy pretrained model, we finetune the same encoder using the combined objective $\mathcal{L}$ defined in Eq> 15, The CS term regularizes the already learned logit space toward a simplex-ETF geometry, while the CE term preserves the original class-discriminative semantics of the logits.

Single-Stage Variant for Graph-Level Detection

To test whether simplex-ETF regularization generalizes beyond node-level detection, we adapt CSafe to graph-level OOD detection, where each input is an independent graph rather than a node embedded in a shared topology. A GNN encoder f_θ with mean or sum graph-level pooling produces a single logit vector $u_i \in \mathbb{R}^C$ per graph G_i, in place of the per-node logits used in the node-level setting. Unlike the two-stage procedure used for node-level CSafe– cross-entropy pretraining followed by a separate CS finetuning stage – we optimize the combined objective from Eq. 15 jointly from initialization in a single training stage, since graph-level classifiers do not require the same warm-start stabilization before geometry regularization is applied.

Inference

Inference is unchanged from energy-based graph OOD detection. Given the trained GNN output u_i, we compute

$$S_i = T \log \sum_{c=1}^{C} \exp(u_{i,c}/T). \quad (17)$$

If propagation is disabled, S_i is used directly as the confidence score. If propagation is enabled, the score is smoothed as

$$\widetilde{S}^{(k+1)} = \eta\widetilde{S}^{(k)} + (1-\eta)P\widetilde{S}^{(k)}. \quad (18)$$

The final score for node i is $\widetilde{S}_i^{(K)}$ after K propagation steps. Since CSafe only modifies training, the test-time detector remains directly comparable to GNNSafe, NodeSafe, and standard energy-based baselines. For graph-level OOD detection, we skip the propagation step and use the raw energy score directly.

The end-to-end training and inference algorithms are detailed in Appendix A

5 Theoretical Analysis

We analyze how the second-stage CSafe finetuning objective improves energy-based node-level OOD detection. The first training stage learns a standard cross-entropy classifier; the second stage regularizes its logit space using CSafe, so the analysis below concerns the CSafe finetuning stage. We defer proofs to Appendix B.

Recall from Section 4 that CS regularizes the GNN output logits $u_i = f_\theta(x_i, \mathcal{G}) \in \mathbb{R}^C$ toward scaled simplex-ETF prototypes $p_c = \rho q_c$ (Eq. 10), where the canonical ETF satisfies the orthogonality property in Eq. 8. This pairwise structure implies the prototypes are uniformly separated:

$$\|p_c - p_k\|_2^2 = \rho^2 \frac{2C}{C-1}, \qquad c \neq k. \quad (19)$$

For a training node with label y_i, CS computes margin-scaled distances $\widetilde{d}_{i,c}$ (Eq. 12) and the corresponding CS logits $z_{i,c}^{\text{CS}} = -s\,\widetilde{d}_{i,c}$ (Eq. 13). At inference, these CS logits are discarded; detection uses only the original GNN logits and the standard negative energy score

$$S(u_i) = T \log \sum_{c=1}^{C} \exp(u_{i,c}/T). \quad (20)$$

The results below characterize how this margin-based distance regularization (i) contracts ID logits toward their assigned prototype, (ii) bounds the resulting deviation in negative energy score, and (iii) propagates through K steps of scalar energy propagation.

Proposition 1 (CS Margin Implies Rigorous Prototype Contraction)**.** *Let $\{p_c\}_{c=1}^C$ be scaled simplex-ETF prototypes with $p_c = \rho q_c$, where $\rho > 0$. Suppose node i with label y_i is correctly classified by the CS distance logits. Then, for every $c \neq y_i$,*

$$e^{m/2}\|u_i - p_{y_i}\|_2 < \|u_i - p_c\|_2.$$

Consequently, for $m > 0$,

$$\|u_i - p_{y_i}\|_2 < \frac{\|p_{y_i} - p_c\|_2}{e^{m/2} - 1} = \frac{\rho}{e^{m/2} - 1}\sqrt{\frac{2C}{C-1}}.$$

Proposition 1 shows that the CSafe margin imposes a geometric contraction around the assigned prototype. Larger m tightens this contraction exponentially. This is different from ordinary cross-entropy, which can separate classes without imposing a fixed reference geometry.

Theorem 1 (Exact ETF Energy and ID Energy Concentration)**.** *Let*

$$S(u) = T \log \sum_{c=1}^{C} \exp(u_c/T)$$

denote the negative energy score. Every scaled canonical simplex-ETF prototype has the same energy

$$S(p_c) = \mu_{\mathrm{ETF}}, \qquad c \in \{1, \dots, C\},$$

where

$$\mu_{\mathrm{ETF}} = T \log \left[\exp\left(\frac{\rho}{T}\sqrt{\frac{C-1}{C}}\right) + (C-1)\exp\left(-\frac{\rho}{T\sqrt{C(C-1)}}\right)\right]. \tag{21}$$

Moreover, if node i is correctly classified by the CS distance logits and $m > 0$, then

$$|S(u_i) - \mu_{\mathrm{ETF}}| < \epsilon_{\mathrm{rep}},$$

where

$$\epsilon_{\mathrm{rep}} := \frac{\rho}{e^{m/2} - 1}\sqrt{\frac{2C}{C-1}}.$$

The theorem gives a class-independent reference energy for the canonical ETF. Unlike an arbitrary collection of class representatives, the ETF prototypes do not merely occupy a narrow energy band. They have exactly equal negative energy. The CS margin controls how far the energy of a correctly classified ID node can deviate from this common value.

We next analyze the explicit score propagation used at inference. Unlike message passing inside the GNN, this step propagates scalar energy scores, not logit vectors.

Theorem 2 (K-Step Scalar Energy Propagation)**.** *Let*

$$M := \eta I + (1-\eta)P,$$

where P is row-stochastic and $\eta \in [0, 1]$. After K propagation steps,

$$\widetilde{S}^{(K)} = M^K S.$$

Assume that every initial ID node $j \in \mathcal{V}_{\mathrm{ID}}$ satisfies

$$|S_j - \mu_{\mathrm{ETF}}| \le \epsilon_{\mathrm{rep}}.$$

Then, for every ID node i,

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \epsilon_{\mathrm{rep}} \sum_{j \in \mathcal{V}_{\mathrm{ID}}} [M^K]_{ij} + \epsilon_{\mathrm{top}}^{(K)}(i),$$

where

$$\epsilon_{\mathrm{top}}^{(K)}(i) := \sum_{j \in \mathcal{V}_{\mathrm{OOD}}} [M^K]_{ij} |S_j - \mu_{\mathrm{ETF}}|.$$

Equivalently, defining

$$\omega_i^{(K)} := \sum_{j \in \mathcal{V}_{\mathrm{OOD}}} [M^K]_{ij},$$

we obtain

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \left(1 - \omega_i^{(K)}\right)\epsilon_{\mathrm{rep}} + \epsilon_{\mathrm{top}}^{(K)}(i).$$

Theorem 2 shows when energy propagation preserves the CSafe-induced ID energy band. If a node is surrounded by score-consistent ID neighbors, propagation does not enlarge the band. If the neighborhood contains OOD or unstable nodes, the propagated score can deviate through the contamination term $\epsilon_{\mathrm{mix}}(i)$. Thus, the main failure mode is not prototype mixing in logit space, but scalar score contamination during graph propagation.

Proposition 1 shows that CS finetuning contracts ID logits toward scaled simplex prototypes. Theorem 1 transfers this contraction to an explicit bound on ID energy deviation. Theorem 2 ties the energy bound to k-step energy propagation, separating the effect of smoothing due to propagation and the deviation from simplex-ETF.

6 Experiments

Experimental Setup

Datasets. We evaluate on eight standard node-level graph OOD benchmarks: Cora, Citeseer, Pubmed, Amazon-Computer, Amazon-Photo, Coauthor-CS, Coauthor-Physics, and arxiv. For each dataset we consider up to three OOD settings: feature perturbation, structure manipulation, and label leave-out, yielding 20 dataset/OOD-type pairs in total. We use the same dataset splits and preprocessing as GNNSafe and NodeSafe. Details are provided in the Appendix C.

Baselines. We compare against MSP (Hendrycks and Gimpel 2016), Mahalanobis (Lee et al. 2018), Energy (Liu et al. 2020), GKDE (Zhao et al. 2020), GPN (Stadler et al. 2021), GNNSafe (Wu et al. 2023a), and NodeSafe (Yang et al. 2024). Details are provided in the Appendix D.

Architecture and training. We primarily report our results on GCN (Kipf and Welling 2016) in the manuscript. The results on the alternate architectures GAT(Veličković et al. 2018), JKNet(Xu et al. 2018), MixHop(Abu-El-Haija et al. 2019), and MLP(Hu et al. 2021) are provided in the Appendix F. The details of the architectures and training hyperparameter settings are in Appendix E.

Metrics. Following standard practice, we report AUROC, and FPR. Higher AUROC indicates better ID/OOD separation; lower FPR indicates stronger OOD rejection.

Main Results

Table 14 reports node-level OOD detection with a GCN backbone on Cora, Citeseer, and Pubmed under Feature, Structure, and Label shift. CSafe achieves the best or second-best AUROC in 8 of 9 settings, leading on all three shifts

Table 1: GCN node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.847±0.005	0.728±0.005	0.913±0.002	0.781±0.005	0.666±0.003	0.889±0.004	0.831±0.005	0.742±0.003	0.860±0.006
	Energy	0.858±0.006	0.711±0.004	*0.917*±0.001	0.793±0.005	0.662±0.004	*0.902*±0.002	0.863±0.010	0.764±0.026	0.863±0.007
	Mahalanobis	0.593±0.018	0.647±0.008	0.314±0.025	0.453±0.016	0.534±0.007	0.249±0.022	0.424±0.020	0.487±0.019	0.401±0.005
	GKDE	0.828±0.012	0.686±0.005	0.572±0.013	0.747±0.012	0.615±0.007	0.827±0.012	0.823±0.014	0.740±0.014	0.834±0.005
	GPN	0.859±0.012	0.775±0.005	0.903±0.013	0.785±0.012	0.706±0.007	0.857±0.012	0.826±0.014	0.750±0.014	0.865±0.005
	GNNSafe	*0.930*±0.005	*0.875*±0.002	0.933±0.001	*0.844*±0.007	*0.805*±0.006	**0.909**±0.002	*0.958*±0.007	*0.926*±0.008	*0.878*±0.004
	NodeSafe	0.964±0.006	0.926±0.003	0.882±0.015	**0.924**±0.020	**0.904**±0.022	0.908±0.007	**0.976**±0.008	0.954±0.013	0.907±0.030
	CSafe	**0.970**±0.003	**0.932**±0.001	**0.934**±0.011	0.907±0.006	0.884±0.007	**0.909**±0.086	0.974±0.02	**0.956**±0.01	**0.919**±0.15
FPR@95	MSP	0.721±0.020	0.881±0.010	0.412±0.025	0.732±0.013	0.862±0.005	0.472±0.033	0.677±0.021	0.828±0.008	0.463±0.017
	Energy	0.642±0.027	0.884±0.014	*0.361*±0.023	0.680±0.026	0.857±0.016	*0.407*±0.015	0.617±0.071	0.768±0.058	0.455±0.032
	Mahalanobis	0.977±0.005	0.912±0.012	0.987±0.007	0.985±0.002	0.937±0.007	0.993±0.002	0.994±0.002	0.989±0.005	0.984±0.005
	GKDE	0.682±0.026	0.843±0.013	0.890±0.012	0.712±0.021	0.937±0.011	0.506±0.010	0.686±0.019	0.815±0.026	0.695±0.013
	GPN	0.562±0.026	0.762±0.013	0.374±0.012	0.731±0.021	0.783±0.011	0.414±0.010	0.618±0.019	0.803±0.026	0.502±0.013
	GNNSafe	*0.459*±0.047	*0.737*±0.014	**0.297**±0.016	*0.646*±0.040	*0.725*±0.014	0.386±0.018	*0.222*±0.036	*0.453*±0.046	**0.379**±0.020
	NodeSafe	0.138±0.030	0.369±0.028	0.626±0.122	0.425±0.180	0.535±0.135	**0.321**±0.020	**0.108**±0.031	0.294±0.147	*0.403*±0.179
	CSafe	**0.091**±0.008	**0.320**±0.01	0.317±0.018	**0.313**±0.05	**0.507**±0.04	0.456±0.18	0.141±0.26	**0.241**±0.15	0.399±0.28

Table 2: GCN node-level graph OOD detection on Amazon, Coauthor, and Arxiv graphs. ID accuracy corresponds to the configuration selected by AUROC. Higher AUROC and ID accuracy are better; lower FPR@95 is better. **Bold**: best; underline: second; *italic*: third.

Dataset	OOD	MSP			Energy			GNNSafe			Mahalanobis			CSafe			NodeSafe		
		AUROC	FPR	Acc.	AUROC	FPR	Acc.	AUROC	FPR	Acc.	AUROC	FPR	Acc.	AUROC	FPR	Acc.	AUROC	FPR	Acc.
Amazon-Computer	Feature	0.709	0.935	0.866	0.815	0.797	**0.879**	0.899	*0.741*	*0.871*	0.319	0.983	0.864	**0.974**	0.033	0.878	**0.974**	**0.008**	0.823
Amazon-Computer	Structure	0.654	*0.999*	0.870	0.730	0.998	*0.871*	*0.969*	**0.000**	**0.878**	0.480	1.000	0.866	**0.983**	**0.000**	0.877	0.977	**0.000**	0.802
Amazon-Photo	Feature	0.954	0.227	0.925	0.972	*0.096*	0.925	*0.984*	0.008	0.925	0.390	0.983	*0.922*	**0.991**	0.008	**0.931**	0.985	**0.004**	0.842
Amazon-Photo	Label	0.937	0.333	0.955	0.918	0.386	**0.958**	0.970	**0.085**	**0.958**	0.477	0.959	0.955	**0.978**	0.091	*0.954*	*0.965*	*0.130*	0.955
Amazon-Photo	Structure	0.989	0.041	0.922	0.965	*0.217*	0.925	*0.987*	**0.000**	0.925	0.560	1.000	*0.923*	**0.993**	**0.000**	**0.927**	0.986	**0.000**	0.895
Coauthor-CS	Feature	0.966	0.185	0.928	0.978	0.096	**0.929**	*0.996*	*0.005*	**0.929**	0.134	0.998	0.928	**1.000**	**0.001**	*0.922*	0.998	0.003	0.919
Coauthor-CS	Label	0.945	0.272	**0.953**	0.957	*0.199*	**0.953**	0.971	0.131	**0.953**	0.382	0.975	**0.953**	**0.973**	**0.113**	0.952	*0.960*	0.207	*0.941*
Coauthor-CS	Structure	0.949	0.268	**0.928**	0.962	0.181	**0.928**	*0.996*	*0.003*	**0.928**	0.392	0.992	**0.928**	**1.000**	**0.000**	0.921	0.997	0.002	*0.919*
Coauthor-Physics	Feature	0.891	0.729	0.958	0.943	0.344	**0.959**	0.993	0.019	**0.959**	0.170	0.999	0.958	**0.999**	**0.003**	*0.957*	*0.986*	*0.038*	0.950
Coauthor-Physics	Structure	0.890	0.803	0.958	0.950	*0.360*	**0.959**	0.998	**0.000**	**0.959**	0.290	1.000	0.958	**1.000**	**0.000**	*0.957*	*0.992*	0.001	0.950
Arxiv	–	0.595	0.926	*0.531*	0.651	0.900	0.533	*0.713*	*0.871*	**0.537**	54.3	95.2	68.32	**0.733**	**0.811**	0.491	0.724	0.825	–

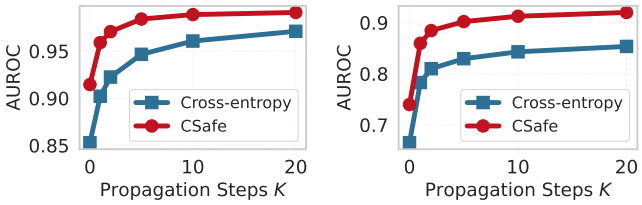
for Cora (0.970, 0.932, 0.934), improving over GNNSafe by 4.0, 5.7, and 0.1 points and surpassing NodeSafe most clearly under label shift (0.882). On Citeseer and Pubmed, NodeSafe is marginally ahead under feature shift (0.924 vs. 0.907; 0.976 vs. 0.974), while CSafe leads on structure and label shift for Pubmed (0.956, 0.919) and ties GNNSafe on Citeseer-Label (0.909). For FPR, CSafe is best or tied-best in 6 of 9 settings, though GNNSafe retains an edge on label shift (Cora: 0.297 vs. 0.317; Pubmed: 0.379 vs. 0.399), and NodeSafe slightly outperforms on Citeseer-Label (0.321 vs. 0.456) and Pubmed-Feature (0.108 vs. 0.141). Table 2 extends the evaluation to Amazon (Computer, Photo), Coauthor (CS, Physics), and the larger-scale Arxiv dataset. CSafe's advantage is more consistent here, achieving the best AUROC in all 9 combinations, including perfect detection on Coauthor-CS (1.000, both shifts) and Coauthor-Physics-Structure (1.000), with gains over GNNSafe ranging from 0.5 points (Amazon-Computer-Structure) to 8.4 points (Amazon-Computer-Feature). CSafe matches or outperforms NodeSafe throughout. On FPR, NodeSafe shows a slight edge on Amazon-Computer-Feature and Amazon-Photo, while several structure-shift cases sit near the floor (0.000) for multiple methods, making AUROC the more informative metric there. On Arxiv, CSafe again leads on both AUROC (0.733) and FPR (0.811), ahead of GNNSafe (0.713, 0.871) and NodeSafe (0.724, 0.825), showing the advantage persists at scale. Overall, CSafe is the strongest method across 16 dataset/shift combinations, with its largest gains under structure shift and its narrowest margins under label shift, where GNNSafe and NodeSafe remain competitive on FPR.

Mechanism Validation

Energy Separation and Stability. Table 3 reports propagated-energy statistics under structure shift, including the variance of propagated ID scores and the mean ID–OOD negative-energy gap. CSafe increases this gap on all three datasets, with an average relative improvement of 29.45%, consistent with our theoretical analysis. Although CSafe constrains correctly classified ID logits around the common ETF energy μ_{ETF}, propagated-score deviations also depend on topology-induced contamination. Thus, the theory does not imply uniformly lower propagated ID variance. Empirically, CSafe reduces ID variance on Citeseer and Cora by an av-

Table 3: Propagated energy statistics under structure shift. Lower ID variance and a larger ID–OOD energy gap are better. The complete results are reported in Appendix G.

	ID Variance ↓		ID–OOD Gap ↑	
Dataset	**CE**	**CSafe**	**CE**	**CSafe**
Citeseer	0.796	**0.699**	0.852	**0.975**
Cora	1.369	**1.242**	1.301	**1.499**
Pubmed	**0.924**	1.698	1.112	**1.775**



(a) Cora-feature OOD (b) Citeseer-structure OOD

Figure 3: After certain number of propagation steps, OOD detection performance saturates but CSafe remains higher.

erage of 10.73%. CSafe substantially enlarges the ID–OOD energy gap on Pubmed. Figure 3 empirically examines the K-dependent behavior characterized by Theorem 2. Increasing K changes the effective propagation weights $[M^K]_{ij}$. It can increase local score consistency, but it can also increase the OOD contamination mass $\omega_i^{(K)}$. The observed saturation is consistent with the competition between these two effects.

Simplex-ETF Logit Geometry Table 4 evaluates complementary aspects of CSafe's logit geometry. Under structure shift, CSafe increases the ETF gap–the difference between the mean OOD and ID distances to the nearest prototype–on Cora (0.609 to 0.738), Citeseer (0.013 to 0.108), and Pubmed (0.292 to 0.309). Thus, OOD logits are more geometrically separated from the simplex prototypes. On Cora label shift, CSafe also reduces the held-out class-mean ETF Gram error from 0.0083 to 0.0059 and slightly improves mean–prototype alignment. However, it does not reduce the normalized sample–prototype distance. These results support stronger class-level simplex-ETF alignment and ID–OOD geometric separation in logit space, rather than exact neural collapse or uniform sample-level collapse.

Ablation Studies

We isolate the contributions of simplex geometry, margin, and cross-entropy in Table 5. Removing the margin tests whether the gain can be explained by centered and rescaled ordinary logits, while the randomly rotated simplex preserves pairwise ETF geometry but changes its orientation. We additionally compare the CS-only objective with the complete CSafe loss. Setting the margin to zero collapses AUROC and FPR back to near-cross-entropy levels (0.902/0.428 vs. 0.901/0.409), showing that the margin, not merely the simplex geometry, drives most of the improvement. Rotating the simplex still outperforms cross-entropy (0.914 AUROC)

Table 4: Simplex-ETF geometry. Panel (a) reports the ETF gap under structure shift; larger is better. Panel (b) reports held-out Cora logit alignment under label shift. Results in Panel (b) are mean ± standard deviation over three seeds.

(a) ETF gap under structure shift		
Dataset	Cross-entropy	CSafe
Cora	0.609	**0.738**
Citeseer	0.013	**0.108**
Pubmed	0.292	**0.309**
(b) Held-out logit alignment on Cora/label		
Metric	Cross-entropy	CSafe
Class-mean ETF Gram MSE ↓	0.0083 ± 0.0023	$\mathbf{0.0059 \pm 0.0001}$
Mean–prototype alignment MSE ↓	0.0022 ± 0.0003	$\mathbf{0.0021 \pm 0.0011}$
Mean–prototype cosine ↑	0.9922 ± 0.0009	$\mathbf{0.9926 \pm 0.0037}$
Normalized sample–prototype distance ↓	$\mathbf{0.5181 \pm 0.0009}$	0.6245 ± 0.0840

Table 5: Mechanism ablation on Cora under label shift. All treatments start from the same seed-matched CE checkpoint. Results are mean ± standard deviation over three seeds.

Method	**AUROC ↑**	**FPR@95 ↓**	**Accuracy ↑**
Cross-entropy	0.901 ± 0.014	0.409 ± 0.038	0.898 ± 0.012
NodeSafe	0.886 ± 0.012	0.503 ± 0.028	$\mathbf{0.910 \pm 0.013}$
CS, $m = 0$	0.902 ± 0.009	0.428 ± 0.038	0.901 ± 0.012
CS, rotated simplex	0.914 ± 0.002	0.379 ± 0.019	0.906 ± 0.001
CS without CE	0.800 ± 0.097	0.682 ± 0.229	0.885 ± 0.015
CSafe	$\mathbf{0.934 \pm 0.011}$	$\mathbf{0.317 \pm 0.018}$	0.897 ± 0.010

but falls short of the full CSafe loss (0.934), indicating that the orientation of the simplex relative to the energy score contributes further gains beyond preserving ETF geometry alone. Dropping cross-entropy entirely degrades both detection performance and stability (0.800 ± 0.097 AUROC), confirming that CE supervision is needed to anchor the representation. Together, these controlled experiments distinguish the contribution of simplex geometry from changes in logit scale, orientation, and supervised loss balance, and show that simplex-ETF geometry, margin, and CE supervision are all necessary for the full improvement of energy-propagation-based node-level OOD detection.

7 Conclusion

In this work, we propose a simplex-ETF based regularizer and two-stage training approach for improved node-OOD detection. The resulting loss is lightweight, keeps inference identical to energy-based propagation, and already improves over strong NODESafe baselines on most benchmarks. The results indicate that, better OOD detection can be achieved by better structuring the logit-space. While we provide limited experiments on graph-level OOD detection and node-OOD detection in heterophilic graphs in the appendix, we will extend CSafe to encompass all types of OOD detection in future work.

References

Abu-El-Haija, S.; Perozzi, B.; Kapoor, A.; Alipourfard, N.; Lerman, K.; Harutyunyan, H.; Ver Steeg, G.; and Galstyan, A. 2019. MixHop: Higher-Order Graph Convolutional Architectures via Sparsified Neighborhood Mixing. In *Pro-*

ceedings of the 36th International Conference on Machine Learning.

Ammar, M. B.; Belkhir, N.; Popescu, S.; Manzanera, A.; and Franchi, G. 2024. NECO: NEural Collapse Based Out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*.

Bao, T.; Wu, Q.; Jiang, Z.; Chen, Y.; Sun, J.; and Yan, J. 2024. Graph Out-of-Distribution Detection Goes Neighborhood Shaping. In *Forty-first International Conference on Machine Learning*.

Bazhenov, G.; Ivanov, S.; Panov, M.; Zaytsev, A.; and Burnaev, E. 2022. Towards OOD Detection in Graph Classification from Uncertainty Estimation Perspective. *arXiv preprint arXiv:2206.10691*.

Bilot, T.; El Madhoun, N.; Al Agha, K.; and Zouaoui, A. 2023. Graph Neural Networks for Intrusion Detection: A Survey. *IEEE Access*, 11: 49114–49139.

Ding, Z.; and Shi, J. 2023. SGOOD: Substructure-Enhanced Graph-Level Out-of-Distribution Detection. *arXiv preprint arXiv:2310.10237*.

Guo, Y.; Yang, C.; Chen, Y.; Liu, J.; Shi, C.; and Du, J. 2023. A Data-Centric Framework to Endow Graph Neural Networks with Out-of-Distribution Detection Ability. *arXiv preprint arXiv:2306.12747*.

Harun, M. Y.; Gallardo, J.; and Kanan, C. 2025. Controlling Neural Collapse Enhances Out-of-Distribution Detection and Transfer Learning. In Singh, A.; Fazel, M.; Hsu, D.; Lacoste-Julien, S.; Berkenkamp, F.; Maharaj, T.; Wagstaff, K.; and Zhu, J., eds., *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 22149–22176. PMLR.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint arXiv:1610.02136*.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *Advances in Neural Information Processing Systems*, 33: 22118–22133.

Hu, Y.; You, H.; Wang, Z.; Wang, Z.; Zhou, E.; and Gao, Y. 2021. Graph-MLP: Node Classification without Message Passing in Graph. arXiv:2106.04051.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations (ICLR 2017)*. Toulon, France.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31.

Li, Z.; Shang, X.; He, R.; Lin, T.; and Wu, C. 2023. No Fear of Classifier Biases: Neural Collapse Inspired Federated Learning with Synthetic and Fixed Classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5319–5329.

Li, Z.; Wu, Q.; Nie, F.; and Yan, J. 2022. GraphDE: A Generative Framework for Debiased Learning and Out-of-Distribution Detection on Graphs. In *Advances in Neural Information Processing Systems*, volume 35, 30277–30290.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing The Reliability of Out-of-Distribution Image Detection in Neural Networks. *arXiv preprint arXiv:1706.02690*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, volume 33, 21464–21475.

Liu, Y.; Ding, K.; Liu, H.; and Pan, S. 2023a. GOOD-D: On Unsupervised Graph Out-of-Distribution Detection. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 339–347.

Liu, Y.; Ding, K.; Lu, Q.; Li, F.; Zhang, L. Y.; and Pan, S. 2023b. Towards self-interpretable graph-level anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36.

Liu, Y.; Ding, K.; Lu, Q.; Li, F.; Zhang, L. Y.; and Pan, S. 2023c. Towards self-interpretable graph-level anomaly detection. *Advances in Neural Information Processing Systems*, 36: 8975–8987.

Ma, J.; Deng, J.; and Mei, Q. 2021. Subgroup generalization and fairness of graph neural networks. *Advances in Neural Information Processing Systems*, 34: 1048–1061.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of Neural Collapse During the Terminal Phase of Deep Learning Training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-GCN: Geometric Graph Convolutional Networks. In *International Conference on Learning Representations*.

Sayyed, A. S.; Bastian, N. D.; and Restuccia, F. 2026. ENCORE: A Neural Collapse Perspective on Out-of-Distribution Detection in Deep Neural Networks. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2944–2953.

Sen, P.; Namata, G. M.; Bilgic, M.; Getoor, L.; Gallagher, B.; and Eliassi-Rad, T. 2008. Collective Classification in Network Data. *AI Magazine*, 29(3): 93–106.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of Graph Neural Network Evaluation. *arXiv preprint arXiv:1811.05868*.

Stadler, M.; Charpentier, B.; Geisler, S.; Z"ugner, D.; and G"unnemann, S. 2021. Graph Posterior Network: Bayesian Predictive Uncertainty for Node Classification. In *Advances in Neural Information Processing Systems*, volume 34, 18033–18048.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Wang, L.; He, D.; Zhang, H.; Liu, Y.; Wang, W.; Pan, S.; Jin, D.; and Chua, T.-S. 2024. Goodat: Towards test-time graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15537–15545.

Wang, Y.; Liu, Y.; Shen, X.; Li, C.; Miao, R.; Ding, K.; Wang, Y.; Pan, S.; and Wang, X. 2025. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark. In *International Conference on Learning Representations*, volume 2025, 91903–91929.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023a. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023b. Energy-based Out-of-Distribution Detection for Graph Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Wu, Q.; Zhang, H.; Yan, J.; and Wipf, D. 2022. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466*.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation Learning on Graphs with Jumping Knowledge Networks. In *Proceedings of the 35th International Conference on Machine Learning*, 5449–5458.

Yang, S.; Liang, B.; Liu, A.; Gui, L.; Yao, X.; and Zhang, X. 2024. Bounded and Uniform Energy-based Out-of-distribution Detection for Graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 56216–56234.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 974–983. London, UK: ACM.

Zhao, X.; Chen, F.; Hu, S.; and Cho, J.-H. 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in neural information processing systems*, 33: 12827–12836.

Zitnik, M.; Agrawal, M.; and Leskovec, J. 2018. Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics*, 34(13): i457–i466.

Supplementary Material

A Algorithm

Algorithm 1: Two-Stage CS Training

Require: In-distribution graph $G = (X, A, Y)$
Require: Training-node indices $\mathcal{I}_{\mathrm{tr}}$, GNN encoder f_θ, number of classes C
Require: Distance-logit scale s, margin m
Require: Loss weights λ_{CE}, λ_{CS}
Require: Learnable unconstrained prototype-radius parameter ρ
Require: Stage-1 epochs E_1, Stage-2 epochs E_2
Ensure: Trained encoder f_θ

Stage 1: Cross-entropy pretraining

for $e = 1, \ldots, E_1$ **do**

Compute GNN logits:

$$Z = f_\theta(X, A), \qquad Z \in \mathbb{R}^{N \times C}. \tag{22}$$

Compute the supervised classification loss:

$$\mathcal{L}^{(1)} = \mathcal{L}_{\mathrm{CE}} = \mathrm{CE}(Z_{\mathcal{I}_{\mathrm{tr}}}, Y_{\mathcal{I}_{\mathrm{tr}}}). \tag{23}$$

Update θ by descending the gradient of $\mathcal{L}^{(1)}$.

end for

Stage 2: CS fine-tuning

Construct canonical simplex-ETF prototypes:

$$Q = \sqrt{\frac{C}{C-1}} \left(I_C - \frac{1}{C} \mathbf{1}\mathbf{1}^\top \right). \tag{24}$$

for $e = 1, \ldots, E_2$ **do**

Compute GNN logits:

$$Z = f_\theta(X, A), \qquad Z \in \mathbb{R}^{N \times C}. \tag{25}$$

Compute the supervised classification loss:

$$\mathcal{L}_{\mathrm{CE}} = \mathrm{CE}(Z_{\mathcal{I}_{\mathrm{tr}}}, Y_{\mathcal{I}_{\mathrm{tr}}}). \tag{26}$$

Compute the positive prototype radius:

$$r = \mathrm{softplus}(\rho). \tag{27}$$

Scale the ETF prototypes:

$$P = rQ. \tag{28}$$

Compute squared distances for all training nodes:

$$D_{ik} = \|Z_i - P_k\|_2^2, \qquad i \in \mathcal{I}_{\mathrm{tr}}, \quad k \in \{1, \ldots, C\}. \tag{29}$$

Apply the target-class margin:

$$D_{i,Y_i} \leftarrow \exp(m) D_{i,Y_i}, \qquad i \in \mathcal{I}_{\mathrm{tr}}. \tag{30}$$

Convert distances into CS logits:

$$H_{ik} = -sD_{ik}. \tag{31}$$

Compute the CS regularization loss:

$$\mathcal{L}_{\mathrm{CS}} = \mathrm{CE}(H, Y_{\mathcal{I}_{\mathrm{tr}}}). \tag{32}$$

Combine the losses:

$$\mathcal{L}^{(2)} = \lambda_{\mathrm{CE}} \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CS}} \mathcal{L}_{\mathrm{CS}}. \tag{33}$$

Update θ and ρ by descending the gradient of $\mathcal{L}^{(2)}$; scale the gradient with respect to ρ by a small constant for stability.

end for

return f_θ

B Proofs

We analyze how the second-stage CS finetuning objective improves energy-based node-level OOD detection. The first training stage learns a standard cross-entropy classifier. The second stage starts from this classifier and regularizes its logit space using CS. Therefore, the analysis below concerns the CS finetuning stage.

Let $u_i = f_\theta(x_i, \mathcal{G}) \in \mathbb{R}^C$ denote the GNN output logits for node i. CS constructs a canonical simplex ETF $\{q_c\}_{c=1}^C$ and scales it by a learnable positive radius $\rho = \mathrm{softplus}(r)$. The class prototype is

$$p_c = \rho q_c. \tag{34}$$

The canonical ETF satisfies

$$q_c^\top q_k = \begin{cases} 1, & c = k, \\ -\dfrac{1}{C-1}, & c \neq k. \end{cases} \tag{35}$$

Hence the scaled prototypes satisfy

$$\|p_c - p_k\|_2^2 = \rho^2 \frac{2C}{C-1}, \qquad c \neq k. \tag{36}$$

For a training node with label y_i, CS computes squared distances

$$d_{i,c} = \|u_i - p_c\|_2^2. \tag{37}$$

The target-class distance is multiplied by the margin factor $\exp(m)$:

$$\widetilde{d}_{i,c} = \begin{cases} \exp(m)\, d_{i,y_i}, & c = y_i, \\ d_{i,c}, & c \neq y_i. \end{cases} \tag{38}$$

The CS logits are then

$$\ell_{i,c}^{\mathrm{CS}} = -\alpha \widetilde{d}_{i,c}, \tag{39}$$

where $\alpha > 0$ is the distance scale. The CS loss applies cross-entropy to these distance logits. At inference, the CS logits are discarded. Detection uses the original GNN logits and the usual negative energy score

$$S(u_i) = T \log \sum_{c=1}^{C} \exp(u_{i,c}/T). \tag{40}$$

Proposition 1(CS Margin Implies Rigorous Prototype Contraction) Let $\{p_c\}_{c=1}^C$ be scaled simplex-ETF prototypes with $p_c = \rho q_c$, where $\rho > 0$. Suppose node i with label y_i is correctly classified by the CS distance logits. Then, for every $c \neq y_i$,

$$e^{m/2} \|u_i - p_{y_i}\|_2 < \|u_i - p_c\|_2.$$

Consequently, for $m > 0$,

$$\|u_i - p_{y_i}\|_2 < \frac{\|p_{y_i} - p_c\|_2}{e^{m/2} - 1} = \frac{\rho}{e^{m/2} - 1} \sqrt{\frac{2C}{C-1}}.$$

Proof. Correct classification by the CS distance logits implies

$$e^m \|u_i - p_{y_i}\|_2^2 < \|u_i - p_c\|_2^2, \qquad c \neq y_i.$$

Taking square roots gives

$$e^{m/2} \|u_i - p_{y_i}\|_2 < \|u_i - p_c\|_2.$$

By the triangle inequality,

$$\|u_i - p_c\|_2 \leq \|u_i - p_{y_i}\|_2 + \|p_{y_i} - p_c\|_2.$$

Therefore,

$$\left(e^{m/2} - 1\right) \|u_i - p_{y_i}\|_2 < \|p_{y_i} - p_c\|_2.$$

For simplex-ETF prototypes,

$$\|p_{y_i} - p_c\|_2 = \rho \sqrt{\frac{2C}{C-1}},$$

which proves the claim. □

Proposition 1 shows that the CS margin imposes a geometric contraction around the assigned prototype. Larger m tightens this contraction exponentially. This is different from ordinary cross-entropy, which can separate classes without imposing a fixed reference geometry.

Theorem 1(Exact ETF Energy and ID Energy Concentration) Let

$$S(u) = T \log \sum_{c=1}^{C} \exp(u_c/T)$$

denote the negative energy score. Every scaled canonical simplex-ETF prototype has the same energy

$$S(p_c) = \mu_{\mathrm{ETF}}, \qquad c \in \{1, \ldots, C\},$$

where

$$\begin{aligned} \mu_{\mathrm{ETF}} = T \log \Bigg[& \exp\left(\frac{\rho}{T}\sqrt{\frac{C-1}{C}}\right) \\ & + (C-1) \exp\left(-\frac{\rho}{T\sqrt{C(C-1)}}\right)\Bigg]. \end{aligned} \tag{41}$$

Moreover, if node i is correctly classified by the CS distance logits and $m > 0$, then

$$|S(u_i) - \mu_{\mathrm{ETF}}| < \epsilon_{\mathrm{rep}},$$

where

$$\epsilon_{\mathrm{rep}} := \frac{\rho}{e^{m/2} - 1} \sqrt{\frac{2C}{C-1}}.$$

Proof. For the canonical simplex ETF,

$$[p_c]_c = \rho\sqrt{\frac{C-1}{C}}, \qquad [p_c]_k = -\frac{\rho}{\sqrt{C(C-1)}}, \quad k \neq c.$$

Substitution into the definition of S yields

$$\begin{aligned} S(p_c) = T \log \Bigg[& \exp\left(\frac{\rho}{T}\sqrt{\frac{C-1}{C}}\right) \\ & + (C-1) \exp\left(-\frac{\rho}{T\sqrt{C(C-1)}}\right)\Bigg]. \end{aligned}$$

This quantity does not depend on c.

Furthermore,

$$\nabla S(u) = \mathrm{softmax}(u/T).$$

Since

$$\|\operatorname{softmax}(u/T)\|_2 \le 1,$$

the function S is 1-Lipschitz under the Euclidean norm. Thus,

$$|S(u_i) - S(p_{y_i})| \le \|u_i - p_{y_i}\|_2.$$

Applying Proposition 1 and using $S(p_{y_i}) = \mu_{\mathrm{ETF}}$ proves the result. □

Theorem 2(K-Step Scalar Energy Propagation) Let

$$M := \eta I + (1-\eta)P,$$

where P is row-stochastic and $\eta \in [0,1]$. After K propagation steps,

$$\widetilde{S}^{(K)} = M^K S.$$

Assume that every initial ID node $j \in \mathcal{V}_{\mathrm{ID}}$ satisfies

$$|S_j - \mu_{\mathrm{ETF}}| \le \epsilon_{\mathrm{rep}}.$$

Then, for every ID node i,

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \epsilon_{\mathrm{rep}} \sum_{j \in \mathcal{V}_{\mathrm{ID}}} [M^K]_{ij} + \epsilon_{\mathrm{top}}^{(K)}(i),$$

where

$$\epsilon_{\mathrm{top}}^{(K)}(i) := \sum_{j \in \mathcal{V}_{\mathrm{OOD}}} [M^K]_{ij} |S_j - \mu_{\mathrm{ETF}}|.$$

Equivalently, defining

$$\omega_i^{(K)} := \sum_{j \in \mathcal{V}_{\mathrm{OOD}}} [M^K]_{ij},$$

we obtain

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \left(1 - \omega_i^{(K)}\right)\epsilon_{\mathrm{rep}} + \epsilon_{\mathrm{top}}^{(K)}(i).$$

Proof. Because P is row-stochastic, M and M^K are row-stochastic. Therefore,

$$\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}} = \sum_j [M^K]_{ij} \left(S_j - \mu_{\mathrm{ETF}}\right).$$

Applying the triangle inequality and partitioning the sum gives

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \sum_{j \in \mathcal{V}_{\mathrm{ID}}} [M^K]_{ij} \left|S_j - \mu_{\mathrm{ETF}}\right| + \sum_{j \in \mathcal{V}_{\mathrm{OOD}}} [M^K]_{ij} \left|S_j - \mu_{\mathrm{ETF}}\right|.$$

The assumed ID concentration bound yields

$$\left|\widetilde{S}_i^{(K)} - \mu_{\mathrm{ETF}}\right| \le \epsilon_{\mathrm{rep}} \sum_{j \in \mathcal{V}_{\mathrm{ID}}} [M^K]_{ij} + \epsilon_{\mathrm{top}}^{(K)}(i).$$

Since each row of M^K sums to one,

$$\sum_{j \in \mathcal{V}_{\mathrm{ID}}} [M^K]_{ij} = 1 - \omega_i^{(K)},$$

which gives the second expression. □

Table 6: Statistics of the node-classification datasets. $|\mathcal{V}|$ and $|\mathcal{E}|$ denote the numbers of nodes and edges, respectively; d denotes the node-feature dimension; and C denotes the number of classes. The edge counts follow the statistics reported by the corresponding PyTorch Geometric or OGB dataset.

Dataset	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	d	C
Cora (Sen et al. 2008)	2,708	10,556	1,433	7
CiteSeer (Sen et al. 2008)	3,327	9,104	3,703	6
PubMed (Sen et al. 2008)	19,717	88,648	500	3
Amazon-Photo (Shchur et al. 2018)	7,650	238,162	745	8
Amazon-Computers (Shchur et al. 2018)	13,752	491,722	767	10
Coauthor-CS (Shchur et al. 2018)	18,333	163,788	6,805	15
Coauthor-Physics (Shchur et al. 2018)	34,493	495,924	8,415	5
Texas (Pei et al. 2020)	183	325	1,703	5
Wisconsin (Pei et al. 2020)	251	515	1,703	5
Cornell (Pei et al. 2020)	183	298	1,703	5
Film (Pei et al. 2020)	7,600	30,019	932	5
ogbn-arxiv (Hu et al. 2020)	169,343	1,166,243	128	40

C Dataset Details

Datasets

We evaluate our method on twelve node-classification benchmarks spanning citation, co-purchase, co-authorship, webpage, social, and temporal citation networks. Table 6 summarizes their principal statistics. The datasets range from small heterophilic graphs containing only a few hundred nodes to the large-scale OGBN-ARXIV network with more than 169,000 nodes. All node features are row-normalized before training, except for OGBN-ARXIV, for which we use the features supplied by OGB.

Citation networks. **Cora**, **CiteSeer**, and **PubMed** are standard citation networks in which nodes represent scientific publications and edges represent citation relationships. Node features are sparse bag-of-words representations of document content, while the labels correspond to research topics. We use the public Planetoid training, validation, and test splits.

Co-purchase and co-authorship networks. **Amazon-Photo** and **Amazon-Computers** are product co-purchase networks. Each node represents a product, an edge connects two products that are frequently purchased together, and the node label denotes the product category. Node features are bag-of-words representations derived from product reviews. **Coauthor-CS** and **Coauthor-Physics** are academic co-authorship networks: nodes represent authors, edges indicate co-authorship, features summarize the keywords of an author's publications, and labels identify the author's primary research area. For datasets without predefined splits, we construct random training, validation, and test partitions using the same proportions for all methods.

Heterophilic networks. **Texas**, **Wisconsin**, and **Cornell** are webpage networks collected from the corresponding university computer-science departments. Nodes denote webpages, edges are hyperlinks, node features are bag-of-words representations of page content, and labels indicate webpage categories. **Film**, also known as the Actor dataset, is an actor co-occurrence network extracted from Wikipedia. Nodes correspond to actors, edges connect actors appearing

on the same Wikipedia page, features encode keywords from the actors' pages, and labels represent five actor categories. These four datasets exhibit relatively low edge homophily and therefore complement the predominantly homophilic citation, co-purchase, and co-authorship benchmarks. We use the first of the ten fixed splits supplied with the PyTorch Geometric versions of these datasets.

Temporal citation network. **ogbn-arxiv** is a directed citation network of computer-science papers from arXiv. Each paper is represented by a 128-dimensional feature vector obtained by averaging word embeddings from its title and abstract, and its label is one of 40 subject areas. Publication years provide a natural source of distribution shift. Following our implementation, papers published up to 2015 form the in-distribution graph, papers from 2016–2017 are used as auxiliary OOD-exposure data, and papers published in 2018, 2019, and 2020 form three temporally ordered OOD test sets. For each temporal partition, we retain only edges whose endpoints are observable by the end of the corresponding period, yielding an inductive evaluation protocol.

OOD construction. For each single-graph benchmark other than OGBN-ARXIV, we consider three complementary forms of distribution shift whenever supported: structural, feature, and label shifts. To construct a *structural shift*, we replace the original adjacency matrix with a graph sampled from a class-conditioned stochastic block model while preserving the nodes, features, and labels. For the homophilic datasets, within-class connections are sampled more frequently than between-class connections. For the heterophilic datasets, this relation is reversed so that the generated graph retains a heterophilic connectivity pattern. A *feature shift* is created by replacing each node feature vector with a random convex combination of the features of two sampled nodes while leaving the graph structure and labels unchanged. For a *label shift*, classes are partitioned into disjoint in-distribution, OOD-exposure, and OOD-test groups. Consequently, OOD nodes belong to classes that are absent from the in-distribution training classes. Label-shift experiments are conducted on Cora, CiteSeer, PubMed, Amazon-Photo, Coauthor-CS, Texas, Wisconsin, Cornell, and Film; Amazon-Computers and Coauthor-Physics are evaluated under structural and feature shifts only.

D Baseline Details

Baseline Methods

We compare against representative post-hoc confidence estimators, distance-based detectors, probabilistic graph models, and graph-aware energy-based approaches. Unless stated otherwise, each baseline uses the same GNN encoder, data splits, and optimization protocol as our method. Higher scores indicate that a node is more likely to be in-distribution (ID).

Maximum Softmax Probability (MSP). MSP (Hendrycks and Gimpel 2016) is a standard post-hoc OOD detector based on the confidence of a classifier trained with cross-entropy. Given the logit vector $\mathbf{z}_v \in \mathbb{R}^C$ of node v, its ID confidence is

$$s_{\mathrm{MSP}}(v) = \max_{c\in\{1,\dots,C\}} \frac{\exp(z_{v,c})}{\sum_{j=1}^{C}\exp(z_{v,j})}. \tag{42}$$

A low maximum probability indicates that the classifier is uncertain and that the node may be OOD. MSP does not require OOD samples during training and does not explicitly use graph structure beyond that already incorporated by the GNN encoder.

Mahalanobis. The Mahalanobis detector (Lee et al. 2018) models the hidden representations of each ID class using a class-conditional Gaussian distribution with a shared covariance matrix. Let $\boldsymbol{\mu}_c$ denote the empirical mean representation of class c and let $\boldsymbol{\Sigma}$ be the shared covariance matrix estimated from the training nodes. The confidence score is

$$s_{\mathrm{Maha}}(v) = \max_{c}\left[-\frac{1}{2}(\mathbf{h}_v-\boldsymbol{\mu}_c)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{h}_v-\boldsymbol{\mu}_c)\right]. \tag{43}$$

Thus, representations that are far from every ID class center receive low ID confidence. In our implementation, the class means and shared precision matrix are estimated from the ID training nodes at the final encoder layer. We additionally tune the magnitude of the optional input perturbation on the validation set.

Energy. Energy-based OOD detection (Liu et al. 2020) assigns each node a scalar score derived from all classifier logits:

$$s_{\mathrm{Energy}}(v) = T\log\sum_{c=1}^{C}\exp\left(\frac{z_{v,c}}{T}\right), \tag{44}$$

where T is the temperature. This quantity is the negative free energy, so larger values represent stronger ID confidence. In contrast to MSP, the energy score retains information about the absolute logit magnitudes rather than depending only on normalized class probabilities. Our Energy baseline uses a classifier trained solely with ID cross-entropy and applies the score independently to each node, without energy propagation or OOD-exposure regularization.

Graph-based Kernel Dirichlet Estimation (GKDE). GKDE (Zhao et al. 2020) is an uncertainty-aware graph model that constructs a Dirichlet distribution over class probabilities at each node. Evidence from labeled nodes is transferred through a graph kernel, with the contribution of a labeled node decreasing as its graph distance from the target node increases. The resulting Dirichlet concentration parameters encode both the predicted class distribution and the amount of available evidence. Nodes receiving little evidence from the labeled ID set have low total concentration and high epistemic uncertainty and are consequently identified as OOD. Unlike MSP and Energy, GKDE explicitly incorporates graph proximity into its uncertainty estimate.

Graph Posterior Network (GPN). GPN (Stadler et al. 2021) combines density-based uncertainty estimation with graph diffusion. It first maps node representations into a latent space and estimates class-conditional densities, commonly using normalizing flows. These densities are converted into

pseudo-counts that parameterize a Dirichlet distribution over class probabilities. The evidence is subsequently propagated over the graph, allowing information from labeled nodes to inform nearby unlabeled nodes. The predictive mean of the Dirichlet distribution is used for node classification, while its total evidence provides an epistemic confidence measure for OOD detection. Nodes lying in low-density regions and receiving limited graph-propagated evidence are assigned high OOD uncertainty.

GNNSafe. GNNSafe (Wu et al. 2023b) extends energy-based detection to graph-structured data by propagating node energies over the graph. Starting from the negative energy $s^{(0)}$, the propagation is

$$\mathbf{s}^{(k+1)} = \eta \mathbf{s}^{(k)} + (1-\eta)\widetilde{\mathbf{A}}\mathbf{s}^{(k)}, \qquad k = 0, \dots, K-1, \tag{45}$$

where $\widetilde{\mathbf{A}}$ is a row-normalized adjacency matrix, K is the number of propagation steps, and η controls the residual contribution of each node's current score. This operation smooths confidence across neighboring nodes and uses local graph context to improve separation between ID and OOD nodes. In our implementation, the GNNSafe classifier is trained using ID cross-entropy only, and the propagated negative energy $\mathbf{s}^{(K)}$ is used as the ID confidence score. No upper-bound regularizer or energy-margin loss is applied.

NodeSafe. NodeSafe augments the graph-propagated energy detector with an upper-bound/uniformity regularizer applied to the classifier logits. For ID training logits $\mathbf{z}_v$, the regularizer controls the dispersion of their norms,

$$\mathcal{L}_{\text{norm}} = \frac{\text{Var}_v\left[\|\mathbf{z}_v\|_2\right]}{\left|\mathbb{E}_v\left[\|\mathbf{z}_v\|_2\right]\right|}, \tag{46}$$

and, after the warm-up period, additionally controls the variance of the logit sums $\sum_c z_{v,c}$. When OOD-exposure samples are available, their logit-sum dispersion is also included in the regularizer. The overall training objective is

$$\mathcal{L}_{\text{NodeSafe}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{UB}}\mathcal{L}_{\text{UB}}, \tag{47}$$

where $\mathcal{L}_{\text{UB}}$ denotes the combined boundedness and uniformity penalty. At inference, NodeSafe computes the negative-energy score and propagates it using the same residual graph diffusion employed by GNNSafe. Therefore, GNNSafe and NodeSafe use the same graph-aware scoring mechanism, but NodeSafe additionally shapes the logit geometry during training to produce more stable and separable energy distributions.

Evaluation protocol. For all methods, OOD detection is evaluated using the ID test nodes as positive samples and the corresponding OOD test nodes as negative samples. We report the area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPR), and the false-positive rate at 95% ID recall (FPR95). Higher AUROC and AUPR are better, whereas lower FPR95 is better. Hyperparameters are selected using validation performance without accessing the OOD test nodes.

Table 7: CSafe architecture, training settings, and hyperparameter search space.

Category	Hyperparameter	Value or search set
Architecture	GNN layers	2
	Hidden channels	64
	Activation	ReLU
	Dropout	0.5
	Batch normalization	Yes
Optimization	Optimizer	Adam
	CE learning rate	0.01
	CS fine-tuning learning rate	0.01
	Weight decay	0.01
	CE training budget	200 epochs
	CS fine-tuning budget	100 epochs
	Selection criterion	Minimum ID validation loss
CS objective	Distance scale s	$\{0.01, 0.1, 1, 2, 4, 8, 16\}$
	Margin m	$\{0.1, 0.25, 0.5, 1, 2, 3, 4\}$
	Radius-gradient multiplier γ_r	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
	CS coefficient λ_{CS}	1
	CE coefficient λ_{CE}	1
	Initial prototype radius	Mean training-logit norm
Detection	Temperature T	1
	Propagation steps K	2
	Residual coefficient η	0.5
	Energy propagation	Enabled

E Implementation and Hyperparameter Settings

Architecture. Unless otherwise stated, we use a two-layer GCN encoder with 64 hidden channels, ReLU activations, batch normalization, and dropout rate 0.5. The final layer produces one logit coordinate per ID class. We use the same architecture and data splits for all methods to ensure a controlled comparison.

Training protocol. CSafe is trained in two stages. We first optimize the encoder using ordinary cross-entropy for at most 200 epochs. Starting from the validation-selected CE checkpoint, we then fine-tune the complete model for at most 100 epochs using the combined objective

$$\mathcal{L} = \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} + \lambda_{\text{CS}}\mathcal{L}_{\text{CS}}.$$

We use $\lambda_{\text{CE}} = \lambda_{\text{CS}} = 1$ in the main experiments. Both stages use Adam with learning rate 0.01 and weight decay 0.01. Within each stage, we retain the checkpoint with the lowest classification loss on the ID validation split.

The simplex radius is initialized using the mean norm of the training logits at the selected CE checkpoint. It remains learnable during CS fine-tuning, with its gradient multiplied by γ_r. Energy scores use temperature $T = 1$ and are propagated for $K = 2$ steps with residual coefficient $\eta = 0.5$.

Hyperparameter selection. We search the CS distance scale s, margin m, and radius-gradient multiplier γ_r over the sets reported in Table 7. For each dataset and OOD-shift pair, the reported CSafe result uses the configuration with the highest AUROC. The broad search uses seed 123; therefore, configuration selection should be interpreted as a descriptive single-seed search rather than a multi-seed estimate of hyperparameter optimality.

Table 8: NodeSafe versus CSafe on heterophilic graphs. Values are mean ± standard deviation over five runs. Bold denotes the better mean within each dataset–OOD pair.

Dataset	OOD type	Method	AUROC ↑	FPR@95 ↓
Cornell	Feature	NodeSafe	0.6552 ± 0.1217	**0.8995 ± 0.0467**
		CSafe	**0.7529 ± 0.0743**	0.9344 ± 0.0479
	Structure	NodeSafe	0.6296 ± 0.1314	0.9410 ± 0.0312
		CSafe	**0.8154 ± 0.0186**	**0.9322 ± 0.0475**
Film	Feature	NodeSafe	0.5477 ± 0.0048	0.9360 ± 0.0088
		CSafe	**0.5758 ± 0.0119**	**0.9060 ± 0.0101**
	Label	NodeSafe	**0.5172 ± 0.0047**	**0.8652 ± 0.0052**
		CSafe	0.4837 ± 0.0306	0.9463 ± 0.0479
	Structure	NodeSafe	0.6220 ± 0.0132	**0.9990 ± 0.0004**
		CSafe	**0.6737 ± 0.0224**	**0.9990 ± 0.0005**
Texas	Feature	NodeSafe	**0.8465 ± 0.0162**	0.9388 ± 0.0174
		CSafe	0.8179 ± 0.0299	**0.8732 ± 0.0296**
	Structure	NodeSafe	0.8228 ± 0.0296	0.9803 ± 0.0175
		CSafe	**0.8612 ± 0.0137**	**0.9301 ± 0.0271**
Wisconsin	Feature	NodeSafe	**0.7155 ± 0.0660**	0.9514 ± 0.0387
		CSafe	0.6837 ± 0.1669	**0.9474 ± 0.0375**
	Structure	NodeSafe	0.7644 ± 0.0580	0.9896 ± 0.0060
		CSafe	**0.8017 ± 0.0230**	**0.9641 ± 0.0232**

NodeSafe and CSafe on Heterophilic Graphs

We compare NodeSafe and CSafe on Cornell, Film, Texas, and Wisconsin under the feature, structure, and available label OOD settings. For each method and dataset–OOD pair, we select the configuration with the highest mean AUROC in its sweep. Results are reported as the mean and standard deviation over five runs. AUROC is better when higher, whereas FPR is better when lower.

CSafe obtains higher mean AUROC in six of the nine settings and lower mean FPR in seven settings. NodeSafe performs better on both metrics for Film under label shift. The methods exhibit a metric trade-off on Cornell under feature shift, Texas under feature shift, and Wisconsin under feature shift, where the method with higher AUROC does not have the lower FPR.

F Additional Results

Tables 9–14 present the results across the different model architectures considered in our evaluation.

G Additional Mechanism Validation

Full Energy Statistic Table 15 presents the full raw and propagated energy statistics for CE and CSafe across all datasets and OOD shift types.

H Performance on Graph-Level OOD Detection

To test whether simplex-ETF regularization transfers beyond node-level detection, we evaluate the single-stage CSafe variant (Section 4) on graph-level OOD benchmarks from UB-GOLD (Wang et al. 2025). We report results using the same train/test splits as UB-GOLD. We compare against four representative graph-level GLAD/GLOD methods from UB-GOLD: GraphDE (Li et al. 2022), SIGNET (Liu et al. 2023c), AAGOD (Guo et al. 2023), and GOODAT (Wang et al. 2024).

Table 16 compares CSafe against four baselines on three UB-GOLD inter-dataset-shift pairs spanning molecule (BZR to COX2, AIDS to DHFR) and protein/enzyme (ENZYMES to PROTEINS) domains, with GraphDE and SIGNET cited from UB-GOLD (Wang et al. 2025) and AAGOD, GOODAT, and a matched CE-only ablation run locally under identical splits and architecture. CSafe improves substantially over the matched CE-only baseline, which shares its architecture and data but omits the CS loss term, across all three pairs (+31.23 on BZR to COX2, +11.51 on AIDS to DHFR, +9.94 on ENZYMES to PROTEINS). Against the matched-supervision baselines AAGOD and GOODAT, CSafe wins clearly on BZR to COX2 (91.37 vs. 77.44 / 80.97) and outperforms AAGOD on ENZYMES to PROTEINS (60.04 vs. 50.17), but trails GOODAT on that pair (64.61) and both baselines narrowly on AIDS to DHFR (93.66 vs. 94.90 / 93.95). Against SIGNET, the strongest fully-unsupervised baseline in UB-GOLD, CSafe leads on BZR to COX2 (91.37 vs. 86.57) and ENZYMES to PROTEINS (60.04 vs. 59.11), but trails on AIDS to DHFR (93.66 vs. 98.84)

I Additional Ablation Results

This appendix provides the complete hyperparameter and mechanism ablations. The main paper reports the controls most directly related to the proposed geometric mechanism.

This appendix provides the complete hyperparameter and mechanism ablations. The main paper reports the controls most directly related to the proposed geometric mechanism.

Margin and Scale Sensitivity

We analyze the sensitivity of CSafe to the margin m and distance scale s on the structure-shift subset. Table 18 reports AUROC and FPR for every tested margin–scale combination.

The effect of the margin is most consistent on Cora, where increasing m generally improves AUROC and reduces FPR. Pubmed also benefits from $m = 1$, particularly at the larger scale $s = 16$. Citeseer is less systematic: its highest AUROC is obtained with $(m, s) = (1, 4)$, but several settings produce similar results, and the configuration with the highest AUROC does not have the lowest FPR. Thus, a positive margin is beneficial for the best-AUROC configurations, whereas sensitivity to the distance scale is dataset-dependent. All configurations use seed 123.

Table 9: GAT node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.896±0.007	0.797±0.004	0.910±0.011	0.840±0.002	0.727±0.004	0.819±0.019	*0.866*±0.009	0.818±0.008	0.892±0.008
	Energy	*0.920*±0.010	*0.801*±0.010	*0.943*±0.008	*0.853*±0.004	0.731±0.007	*0.878*±0.016	0.861±0.032	*0.820*±0.031	*0.891*±0.007
	Mahalanobis	0.344±0.013	0.573±0.006	0.362±0.019	0.377±0.027	0.556±0.036	0.386±0.015	0.356±0.015	0.373±0.015	0.456±0.051
	GNNSafe	0.955±0.007	0.897±0.013	**0.958**±0.005	0.879±0.005	0.847±0.008	0.882±0.012	0.917±0.033	0.885±0.036	0.885±0.007
	NodeSafe	0.794±0.061	0.763±0.052	0.928±0.016	0.732±0.057	*0.736*±0.054	*0.878*±0.016	0.789±0.130	0.772±0.125	**0.901**±0.025
	CSafe	**0.964**±0.005	**0.915**±0.008	0.948±0.005	**0.890**±0.006	**0.859**±0.009	**0.889**±0.005	**0.959**±0.006	**0.931**±0.011	**0.901**±0.016
FPR@95	MSP	0.540±0.044	0.736±0.019	0.520±0.061	0.614±0.018	*0.797*±0.021	0.768±0.058	*0.589*±0.048	*0.706*±0.026	**0.289**±0.009
	Energy	*0.377*±0.086	*0.681*±0.061	*0.281*±0.054	*0.633*±0.060	0.804±0.038	0.422±0.070	0.659±0.166	0.726±0.101	**0.289**±0.016
	Mahalanobis	0.987±0.003	0.927±0.005	0.982±0.004	0.990±0.003	0.897±0.027	0.984±0.005	0.996±0.001	0.991±0.001	0.966±0.016
	GNNSafe	0.207±0.061	0.500±0.117	**0.188**±0.051	0.614±0.089	0.692±0.037	*0.425*±0.064	0.456±0.233	0.602±0.160	0.311±0.012
	NodeSafe	0.923±0.081	0.845±0.059	0.355±0.105	0.946±0.037	0.803±0.026	**0.402**±0.060	0.827±0.240	0.836±0.158	*0.351*±0.055
	CSafe	**0.158**±0.032	**0.447**±0.069	0.228±0.042	**0.543**±0.059	**0.638**±0.047	*0.425*±0.056	**0.195**±0.037	**0.353**±0.065	0.353±0.074

Table 10: GATJK node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.817±0.004	0.705±0.005	0.883±0.002	0.776±0.004	0.652±0.004	*0.889*±0.001	0.780±0.002	0.760±0.002	**0.867**±0.003
	Energy	*0.846*±0.003	0.717±0.004	*0.901*±0.002	0.815±0.003	0.674±0.003	0.909±0.001	0.809±0.002	0.789±0.005	0.863±0.005
	GNNSafe	0.925±0.001	0.839±0.002	**0.926**±0.002	*0.864*±0.002	*0.796*±0.003	**0.912**±0.000	0.934±0.003	0.875±0.007	**0.867**±0.004
	NodeSafe	0.762±0.165	*0.768*±0.151	0.717±0.116	0.880±0.034	0.847±0.035	0.862±0.068	*0.843*±0.172	*0.798*±0.162	*0.857*±0.034
	CSafe	**0.963**±0.003	**0.903**±0.005	0.922±0.003	**0.908**±0.005	**0.858**±0.005	0.884±0.010	**0.968**±0.004	**0.929**±0.007	0.751±0.077
FPR@95	MSP	0.712±0.009	0.856±0.005	0.546±0.028	0.737±0.010	0.864±0.007	*0.479*±0.018	0.713±0.008	0.774±0.006	0.413±0.012
	Energy	*0.695*±0.016	0.866±0.009	*0.442*±0.013	*0.643*±0.015	0.824±0.011	0.380±0.009	0.680±0.012	*0.746*±0.017	*0.427*±0.007
	GNNSafe	0.393±0.025	*0.767*±0.009	**0.311**±0.015	0.584±0.007	*0.740*±0.012	**0.330**±0.013	0.320±0.021	0.606±0.040	**0.381**±0.004
	NodeSafe	0.730±0.350	0.689±0.241	0.845±0.247	0.685±0.235	0.683±0.082	0.548±0.184	*0.578*±0.335	0.749±0.283	0.625±0.242
	CSafe	**0.169**±0.011	**0.525**±0.029	0.402±0.029	**0.430**±0.051	**0.629**±0.024	0.524±0.064	**0.147**±0.016	**0.345**±0.057	0.739±0.223

Table 11: GCNJK node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.723±0.004	0.592±0.003	0.866±0.005	0.708±0.004	0.583±0.003	0.881±0.001	0.707±0.005	0.635±0.006	*0.841*±0.005
	Energy	0.758±0.006	0.598±0.004	*0.877*±0.004	0.746±0.003	0.598±0.001	*0.895*±0.003	0.719±0.013	0.654±0.013	0.833±0.007
	GNNSafe	0.898±0.004	*0.812*±0.004	**0.911**±0.003	*0.826*±0.003	*0.751*±0.002	**0.907**±0.003	*0.906*±0.014	*0.859*±0.009	**0.865**±0.004
	NodeSafe	*0.893*±0.021	0.854±0.019	0.844±0.064	0.840±0.085	0.807±0.037	0.899±0.008	0.955±0.025	**0.956**±0.011	0.836±0.076
	CSafe	**0.951**±0.006	**0.886**±0.004	0.894±0.013	**0.869**±0.013	**0.820**±0.007	0.885±0.016	**0.958**±0.008	0.941±0.014	0.854±0.027
FPR@95	MSP	0.886±0.009	0.952±0.004	0.628±0.041	0.822±0.012	0.915±0.006	0.492±0.016	0.886±0.010	0.937±0.007	*0.550*±0.036
	Energy	0.774±0.019	0.919±0.007	0.553±0.025	0.727±0.017	0.884±0.005	*0.453*±0.015	0.850±0.031	0.888±0.022	0.587±0.044
	GNNSafe	0.617±0.034	*0.821*±0.014	**0.427**±0.014	**0.686**±0.020	*0.776*±0.006	**0.347**±0.045	*0.572*±0.120	*0.790*±0.069	**0.432**±0.029
	NodeSafe	*0.659*±0.140	0.725±0.061	0.658±0.209	*0.755*±0.154	**0.768**±0.034	0.362±0.059	0.241±0.161	**0.238**±0.061	0.651±0.291
	CSafe	**0.241**±0.058	**0.582**±0.036	*0.559*±0.097	0.784±0.105	0.773±0.020	0.455±0.103	**0.233**±0.056	0.344±0.104	0.525±0.151

Table 12: MixHop node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.751±0.022	0.645±0.014	*0.861*±0.016	0.700±0.019	0.601±0.013	0.813±0.019	0.660±0.010	0.605±0.008	**0.813**±0.013
	Energy	0.821±0.006	0.687±0.005	**0.899**±0.006	0.752±0.011	0.634±0.009	**0.865**±0.008	0.732±0.021	0.697±0.018	*0.745*±0.014
	GNNSafe	*0.904*±0.007	*0.873*±0.006	0.880±0.030	*0.809*±0.017	*0.771*±0.012	0.848±0.018	0.894±0.030	*0.879*±0.030	0.725±0.025
	NodeSafe	0.915±0.022	0.932±0.036	0.761±0.066	0.840±0.033	0.843±0.012	*0.825*±0.021	*0.893*±0.046	0.890±0.114	0.777±0.009
	CSafe	**0.965**±0.004	**0.938**±0.003	0.817±0.039	**0.908**±0.007	**0.888**±0.005	0.815±0.033	**0.964**±0.007	**0.967**±0.007	0.724±0.031
FPR@95	MSP	0.812±0.018	0.886±0.008	*0.657*±0.048	0.836±0.020	0.902±0.009	0.724±0.050	0.891±0.013	0.912±0.008	**0.631**±0.034
	Energy	0.714±0.048	0.861±0.018	**0.431**±0.046	*0.793*±0.018	0.887±0.014	**0.555**±0.035	0.832±0.034	0.831±0.023	0.725±0.023
	GNNSafe	*0.556*±0.013	*0.642*±0.031	0.490±0.116	0.846±0.036	*0.781*±0.016	0.625±0.045	0.578±0.151	*0.631*±0.156	0.814±0.025
	NodeSafe	0.499±0.136	0.355±0.196	0.778±0.087	0.655±0.158	0.716±0.037	0.728±0.021	*0.621*±0.205	0.511±0.424	*0.792*±0.046
	CSafe	**0.156**±0.022	**0.304**±0.027	0.676±0.071	**0.482**±0.044	**0.585**±0.041	*0.713*±0.061	**0.202**±0.060	**0.173**±0.039	0.799±0.026

Table 13: MLP node-level graph OOD detection on citation graphs. Higher AUROC and lower FPR are better. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**, underline, and *italic* represent the best, second-best, and third-best method.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.552±0.006	0.477±0.003	0.712±0.013	0.565±0.015	0.487±0.005	0.764±0.009	0.589±0.004	**0.518**±0.004	0.771±0.013
	Energy	0.586±0.003	0.477±0.003	0.743±0.008	0.602±0.007	0.482±0.003	0.786±0.011	0.582±0.004	*0.504*±0.001	0.758±0.009
	Mahalanobis	0.494±0.010	0.476±0.001	0.415±0.031	0.529±0.019	0.482±0.007	0.324±0.010	0.502±0.023	0.499±0.008	0.404±0.034
	GNNSafe	*0.671*±0.006	*0.525*±0.007	0.888±0.005	*0.646*±0.013	*0.550*±0.005	*0.866*±0.008	*0.601*±0.012	0.450±0.008	*0.815*±0.008
	NodeSafe	0.696±0.005	**0.535**±0.003	*0.884*±0.011	0.684±0.016	**0.575**±0.012	0.895±0.017	0.671±0.036	0.506±0.041	**0.924**±0.032
	CSafe	**0.705**±0.007	0.532±0.002	**0.897**±0.006	**0.711**±0.008	0.566±0.005	**0.903**±0.008	**0.692**±0.009	0.500±0.010	0.902±0.015
FPR@95	MSP	0.921±0.010	0.950±0.005	0.874±0.043	0.916±0.008	0.948±0.003	0.746±0.031	**0.930**±0.004	**0.947**±0.005	0.740±0.045
	Energy	0.920±0.010	0.952±0.001	0.826±0.046	**0.896**±0.010	0.947±0.003	0.740±0.030	*0.944*±0.009	0.949±0.005	0.793±0.025
	Mahalanobis	0.965±0.005	0.951±0.004	0.975±0.004	0.951±0.011	0.952±0.006	0.994±0.004	0.969±0.011	*0.954*±0.006	0.965±0.012
	GNNSafe	*0.894*±0.014	0.880±0.007	*0.508*±0.049	*0.914*±0.010	**0.818**±0.002	*0.549*±0.063	0.971±0.006	0.972±0.001	*0.607*±0.040
	NodeSafe	0.885±0.010	*0.883*±0.007	0.485±0.056	0.933±0.018	*0.824*±0.004	0.401±0.051	0.957±0.017	0.972±0.001	**0.300**±0.073
	CSafe	**0.865**±0.012	**0.878**±0.003	**0.418**±0.031	0.912±0.011	0.823±0.002	**0.349**±0.033	0.939±0.005	0.972±0.001	0.380±0.049

Table 14: GCN node-level graph OOD detection on citation graphs. Higher AUROC and ID accuracy are better; lower FPR@95 is better. ID accuracy corresponds to the configuration selected by AUROC. Results are reported as mean ± standard deviation when multiple runs are available; values without an uncertainty term correspond to single-run experiments. **Bold**: best; underline: second; *italic*: third.

Metric	Method	Cora			Citeseer			Pubmed		
		Feature	Structure	Label	Feature	Structure	Label	Feature	Structure	Label
AUROC	MSP	0.847±0.005	0.728±0.005	0.913±0.002	0.781±0.005	0.666±0.003	0.889±0.004	0.831±0.005	0.742±0.003	0.860±0.006
	Energy	0.858±0.006	0.711±0.004	*0.917*±0.001	0.793±0.005	0.662±0.004	*0.902*±0.002	0.863±0.010	0.764±0.026	0.863±0.007
	Mahalanobis	0.593±0.018	0.647±0.008	0.314±0.025	0.453±0.016	0.534±0.007	0.249±0.022	0.424±0.020	0.487±0.019	0.401±0.005
	GKDE	0.828±0.012	0.686±0.005	0.572±0.013	0.747±0.012	0.615±0.007	0.827±0.012	0.823±0.014	0.740±0.014	0.834±0.005
	GPN	0.859±0.012	0.775±0.005	0.903±0.013	0.785±0.012	0.706±0.007	0.857±0.012	0.826±0.014	0.750±0.014	0.865±0.005
	GNNSafe	*0.930*±0.005	*0.875*±0.002	0.933±0.001	*0.844*±0.007	*0.805*±0.006	**0.909**±0.002	*0.958*±0.007	*0.926*±0.008	*0.878*±0.004
	NodeSafe	0.964±0.006	0.926±0.003	0.882±0.015	**0.924**±0.020	**0.904**±0.022	0.908±0.007	**0.976**±0.008	0.954±0.013	0.907±0.030
	CSafe	**0.970**±0.003	**0.932**±0.001	**0.934**±0.011	0.907±0.006	0.884±0.007	**0.909**±0.086	0.974±0.02	**0.956**±0.01	**0.919**±0.15
FPR@95	MSP	0.721±0.020	0.881±0.010	0.412±0.025	0.732±0.013	0.862±0.005	0.472±0.033	0.677±0.021	0.828±0.008	0.463±0.017
	Energy	0.642±0.027	0.884±0.014	*0.361*±0.023	0.680±0.026	0.857±0.016	*0.407*±0.015	0.617±0.071	0.768±0.058	0.455±0.032
	Mahalanobis	0.977±0.005	0.912±0.012	0.987±0.007	0.985±0.002	0.937±0.007	0.993±0.002	0.994±0.002	0.989±0.005	0.984±0.005
	GKDE	0.682±0.026	0.843±0.013	0.890±0.012	0.712±0.021	0.937±0.011	0.506±0.010	0.686±0.019	0.815±0.026	0.695±0.013
	GPN	0.562±0.026	0.762±0.013	0.374±0.012	0.731±0.021	0.783±0.011	0.414±0.010	0.618±0.019	0.803±0.026	0.502±0.013
	GNNSafe	*0.459*±0.047	*0.737*±0.014	**0.297**±0.016	*0.646*±0.040	*0.725*±0.014	0.386±0.018	*0.222*±0.036	*0.453*±0.046	**0.379**±0.020
	NodeSafe	0.138±0.030	0.369±0.028	0.626±0.122	0.425±0.180	0.535±0.135	**0.321**±0.020	**0.108**±0.031	0.294±0.147	*0.403*±0.179
	CSafe	**0.091**±0.008	**0.320**±0.01	0.317±0.018	**0.313**±0.05	**0.507**±0.04	0.456±0.18	0.141±0.26	**0.241**±0.15	0.399±0.28
ID Acc.	MSP	*0.775*±0.009	0.776±0.009	*0.898*±0.006	*0.656*±0.007	*0.656*±0.007	*0.896*±0.005	0.752±0.004	0.752±0.004	**1.000**±0.000
	Energy	0.772±0.004	0.773±0.004	0.901±0.004	0.655±0.008	*0.656*±0.006	0.899±0.005	*0.765*±0.008	0.762±0.009	**1.000**±0.000
	Mahalanobis	*0.776*±0.009	0.776±0.009	*0.898*±0.006	*0.656*±0.007	*0.656*±0.007	*0.896*±0.005	0.752±0.004	0.752±0.004	**1.000**±0.000
	GKDE	0.748±0.002	0.737±0.006	0.898±0.005	0.642±0.001	0.647±0.008	0.8916±0.010	0.741±0.001	0.752±0.001	1.00±0.000
	GPN	0.771±0.003	0.764±0.006	0.914±0.009	0.631±0.005	0.658±0.001	0.893±0.006	0.758±0.001	0.745±0.000	1.00±0.00
	GNNSafe	0.772±0.004	0.773±0.005	0.901±0.004	0.655±0.007	0.655±0.008	0.899±0.005	0.766±0.008	*0.760*±0.008	**1.000**±0.000
	NodeSafe	0.785±0.005	**0.792**±0.004	0.901±0.006	**0.685**±0.017	**0.684**±0.018	**0.906**±0.008	**0.779**±0.009	**0.777**±0.008	**1.000**±0.000
	CSafe	**0.792**±0.003	*0.774*±0.004	**0.915**±0.002	0.664±0.004	0.664±0.005	0.891±0.002	0.741±0.004	0.753±0.000	**1.000**

Table 15: Energy statistics for the cross-entropy baseline (CE) and CSafe. Lower ID variance indicates greater within-ID concentration, while a larger ID–OOD energy gap indicates stronger score separation.

Dataset	OOD	CE Var_{raw}	CSafe Var_{raw}	CE Var_{prop}	CSafe Var_{prop}	CE Gap_{raw}	CSafe Gap_{raw}	CE Gap_{prop}	CSafe Gap_{prop}
Citeseer	Feature	0.770	0.735	0.911	0.842	0.566	0.696	0.784	0.920
Citeseer	Label	1.264	2.214	1.457	2.163	1.172	1.708	1.404	1.947
Citeseer	Structure	0.666	0.609	0.796	0.699	0.298	0.408	0.852	0.975
Cora	Feature	1.439	1.328	1.770	1.496	0.992	1.135	1.523	1.674
Cora	Label	1.659	2.037	1.403	1.656	1.356	1.520	1.550	1.644
Cora	Structure	1.146	1.037	1.369	1.242	0.562	0.703	1.301	1.499
Pubmed	Feature	0.323	1.286	0.382	1.773	0.437	0.783	0.774	1.414
Pubmed	Label	1.600	7.897	2.345	5.671	0.946	1.988	1.390	2.945
Pubmed	Structure	0.810	1.625	0.924	1.698	0.547	0.925	1.112	1.775

Table 16: Graph-level OOD detection results.

Dataset	GraphDE	SIGNET	AAGOD	GOODAT	CE-only	CSafe (ours)
BZR+COX	56.34	86.57	77.44	80.97	60.14	91.37
AIDS+DHFR	94.58	98.84	94.90	93.95	82.15	93.66
ENZYMES+ PROTEINS	54.48	59.11	50.17	64.61	50.1	60.04

Table 17: Complete mechanism ablation on Cora under label shift. All treatments start from the same seed-matched, validation-selected CE checkpoint and receive the same additional optimization budget. Results are mean ± standard deviation over three seeds.

Treatment	**AUROC** ↑	**AUPR** ↑	**FPR@95** ↓	**Accuracy** ↑
Cross-entropy	0.901 ± 0.014	0.752 ± 0.021	0.409 ± 0.038	0.898 ± 0.012
CE + logit ℓ_2	0.914 ± 0.002	$\mathbf{0.785 \pm 0.005}$	0.379 ± 0.019	0.906 ± 0.001
CE + logit centering	0.911 ± 0.003	0.773 ± 0.003	0.374 ± 0.015	0.901 ± 0.015
CE + learned center loss	0.865 ± 0.010	0.658 ± 0.025	0.511 ± 0.007	0.902 ± 0.004
CS, fixed one-hot prototypes	0.907 ± 0.006	0.769 ± 0.015	0.396 ± 0.011	0.905 ± 0.005
CS, randomly rotated simplex	0.914 ± 0.002	0.785 ± 0.005	0.379 ± 0.019	0.906 ± 0.001
CS, fixed simplex radius	$\mathbf{0.934 \pm 0.011}$	0.782 ± 0.011	$\mathbf{0.317 \pm 0.018}$	0.907 ± 0.004
CS, learned simplex radius (CSafe)	0.903 ± 0.011	0.760 ± 0.022	0.405 ± 0.018	0.897 ± 0.010
CS, $m = 0$	0.902 ± 0.009	0.752 ± 0.022	0.428 ± 0.038	0.901 ± 0.012
CS without CE	0.800 ± 0.097	0.611 ± 0.153	0.682 ± 0.229	0.885 ± 0.015
NodeSafe	0.886 ± 0.012	0.728 ± 0.025	0.503 ± 0.028	$\mathbf{0.910 \pm 0.013}$

Table 18: Sensitivity of FEO-Softmax to the margin m and distance scale s on the structure-shift subset using seed 123. Each entry reports AUROC/FPR@95. Bold entries identify the highest-AUROC configuration for each dataset.

Dataset	m	$s = 4$	$s = 8$	$s = 16$
Citeseer	0.0	0.821/0.617	0.849/0.574	0.828/0.636
	0.1	0.833/0.646	0.831/0.611	0.820/0.673
	0.5	0.772/0.750	0.854/0.560	0.796/0.718
	1.0	**0.872/0.656**	0.860/0.677	0.844/0.657
Cora	0.0	0.842/0.782	0.893/0.663	0.892/0.557
	0.1	0.874/0.681	0.889/0.589	0.894/0.504
	0.5	0.903/0.506	0.902/0.508	0.901/0.527
	1.0	0.921/0.404	0.915/0.476	**0.926/0.369**
Pubmed	0.0	0.942/0.298	0.921/0.408	0.895/0.611
	0.1	0.935/0.387	0.933/0.425	0.915/0.589
	0.5	0.900/0.650	0.933/0.389	0.936/0.358
	1.0	0.937/0.383	0.942/0.350	**0.950/0.242**