STRENGTHENING UNSUPERVISED GRAPH OUT-OF-DISTRIBUTION DETECTION WITH STRUCTURAL RESONANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing graph Out-of-distribution (OOD) detection methods rely on in-distribution (ID) category labels to obtain meaningful representations, which limits their applicability in unlabeled settings. Recent representative unsupervised graph OOD detection methods exploit ID–OOD discrepancy in optimization-induced representation dynamics (Resonance), mitigating the dependence on labels. However, these methods still fall short in fully leveraging the intrinsic structural properties of graphs. Moreover, resonance inferred from single-period optimization-induced changes are highly sensitive to stochastic noise and transient gradient effects, which undermines estimation reliability. To address these challenges, we propose **Multi-Scale Structural Resonance Estimation (MSRE)**, which models local structural resonance and performs multi-scale temporal aggregation, and we further provide a consistency-based candidate selection strategy that requires no labeled validation set. Extensive experiments on multiple benchmark graphs demonstrate that the proposed framework consistently outperforms state-of-the-art unsupervised OOD detection methods.

1 INTRODUCTION

Graph-structured data are ubiquitous in a wide spectrum of real-world applications, including financial fraud detection, recommender systems, molecular property analysis, and cybersecurity monitoring Zhou et al. (2020). In such scenarios, detecting out-of-distribution (OOD) nodes or instances is crucial for ensuring system reliability and trustworthy deployment. This problem, referred to as graph OOD detection, is particularly challenging due to the intrinsic complexity of graph data, such as multi-relational interactions, semantic heterogeneity, and potentially low homophily Pei et al. (2020). Although graph neural networks (GNNs) Kipf & Welling (2017); Veličković et al. (2018) have achieved remarkable success on numerous graph learning tasks, they remain sensitive to distribution shifts, necessitating effective OOD detection for graph-structured data.

A straightforward way to adapt conventional OOD detection techniques to graph domains is to use classifier-derived confidence or energy scores Hendrycks & Gimpel (2017); Liu et al. (2020). However, these scoring rules do not themselves explicitly model the relational dependencies among nodes. Graph-specific approaches incorporate structural information at different task levels. GraphDE Li et al. (2022) develops a generative framework for debiased graph learning and graph-level OOD detection, whereas OODGAT Song & Wang (2022) and GNNSafe Wu et al. (2023) address node-level OOD detection through attention-based message passing and energy-based scoring with propagation, respectively. Many such approaches build on ID category supervision, motivating methods that can operate without semantic class labels.

To alleviate the reliance on labeled ID data, recent studies have explored graph anomaly and OOD detection without category supervision Ding et al. (2019); Gong & Sun (2024). Among them, RSL Yang et al. (2025) is a representative approach that leverages feature resonance to characterize OOD nodes. It aligns the representations of known ID nodes with arbitrary targets and quantifies resonance by measuring the representation changes induced by this alignment. This resonance-based sensitivity shifts the focus of OOD detection from output uncertainty to representation dynamics, providing a detection signal without requiring semantic class labels.

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.

Despite their intuitive motivation, existing resonance-based graph OOD detection methods exhibit several limitations. **First**, existing resonance-based methods primarily exploit node feature resonance to alleviate the reliance on label information. However, such feature-only measurements fail to fully leverage the intrinsic structural properties of graphs, as they do not explicitly characterize how optimization-induced perturbations propagate and interact through multi-hop neighborhood aggregation. Although GNN representations already incorporate graph topology, measuring the response magnitude at each node does not fully capture the collective structural responses of neighboring nodes. **Second**, estimating resonance signals solely from feature alignment variations between adjacent training epochs is highly susceptible to stochastic optimization noise and transient gradient fluctuations, resulting in unstable resonance estimation across training stages. **Third**, existing resonance-based methods rely on validation sets to identify resonance-critical training stages, which is impractical when labeled ID/OOD validation data are unavailable. This dependence limits their applicability in settings that require both category-label-free learning and validation-free candidate selection.

Beyond node-wise feature dynamics, graph topology determines how optimization-induced representation perturbations propagate and interact across local neighborhoods, providing an additional source of OOD evidence. Resonance is therefore jointly shaped by representation dynamics and graph structure, motivating the modeling of both individual and collective responses across training stages. We propose **Multi-Scale Structural Resonance Estimation (MSRE)**, an unsupervised framework that explicitly leverages multi-scale structural resonance for stable, validation-free OOD identification. MSRE models the propagation of inter-epoch representation perturbations through multi-hop neighborhoods. By aggregating these perturbation vectors before computing their magnitude, it explicitly preserves directional interactions among neighboring perturbations during propagation. Aligned perturbations reinforce one another, whereas oppositely directed perturbations can partially cancel, producing topology-conditioned responses that cannot be recovered from node-wise perturbation magnitudes alone. MSRE combines these structural responses with feature resonance and aggregates the joint signals across training stages to suppress stochastic optimization noise and gradient fluctuations. A validation-free resonance consensus strategy then identifies persistently weak-resonance nodes as OOD candidates, eliminating the need for labeled validation data in resonance-stage selection.

Our contributions are summarized as follows:

- We introduce structural resonance as a complementary OOD signal to feature resonance by explicitly modeling the propagation of optimization-induced representation perturbations through graph neighborhoods. It preserves directional interactions among inter-epoch representation perturbations during neighborhood propagation, capturing topology-weighted reinforcement and cancellation beyond individual response magnitudes. This enables resonance-based OOD characterization to more fully exploit the intrinsic structural information of graphs.
- We propose a multi-scale temporal resonance aggregation scheme that stabilizes resonance estimation across training stages, suppressing stochastic optimization noise and transient gradient fluctuations associated with single-period measurements.
- We design a validation-free resonance consensus strategy for candidate selection, which identifies reliable OOD candidates based on temporally consistent weak-resonance patterns, eliminating the need for labeled validation data in resonance-stage selection.
- Through experiments on five graph benchmarks, we demonstrate competitive OOD detection performance and improvements over RSL under a matched validation-assisted protocol. Additional comparisons evaluate the validation-free version, while targeted ablations examine structural directional interactions and the contributions of the other components.

2 PRELIMINARIES

2.1 LEARNING SETUP

We consider node-level out-of-distribution (OOD) detection on a graph under an unsupervised setting. Let $\mathcal{G} = (V, E, X)$, where $V = \{v_1, \ldots, v_{|V|}\}$ is the node set, $E \subseteq V \times V$ is the edge set,

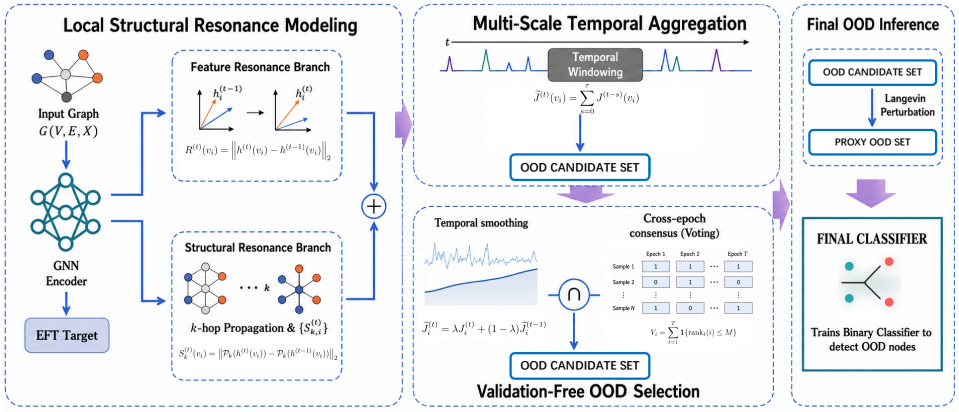


Figure 1: Overview of the proposed MSRE framework.

and $X \in \mathbb{R}^{|V|\times d}$ is the node feature matrix. A graph neural network (GNN) with parameters θ performs message passing on $\mathcal{G}$ and yields node representations over training epochs; we denote the representation of node v_i at epoch t by $h_i^{(t)} \in \mathbb{R}^d$.

Unlike supervised or semi-supervised OOD detection, we have access to neither semantic class labels nor a pretext classification objective, and we assume no labeled validation set. Instead, we are given a small set of ID nodes $V_{\text{ID}} \subset V$ that are known to be in-distribution, without category annotations. These seeds are used only to induce controlled perturbations during training and are not tied to any explicit prediction objective. Our objective is to identify a subset $V_{\text{OOD}} \subset V$ whose nodes exhibit feature and/or structural patterns that deviate from the dominant in-distribution population, using only the training dynamics of the GNN.

3 METHOD

To address the noise sensitivity of single-period resonance and the lack of structural modeling, we propose **Multi-Scale Structural Resonance Estimation (MSRE)**. MSRE (i) models local structural resonance, (ii) performs multi-scale temporal aggregation, and (iii) adopts a validation-free consistency-based candidate selection strategy to identify reliable OOD candidates without labeled validation data.

3.1 LOCAL STRUCTURAL RESONANCE MODELING

Feature resonance refers to the phenomenon that unknown ID nodes typically exhibit stronger optimization-induced representation changes than OOD nodes when a GNN is trained with objectives defined on known ID nodes. A detailed introduction is provided in Appendix A.4. However, in GNNs, node representations are not independent: they are generated through neighborhood aggregation and are inherently coupled by graph topology. Furthermore, although GNN representations already encode graph structure through message passing, measuring the change magnitude at each node does not explicitly retain the directional relations among neighboring representation changes. We introduce *structural resonance* to capture these relations by propagating inter-epoch representation perturbations through local topology before measuring their magnitude.

Micro-level structural resonance. Let $H^{(t)} \in \mathbb{R}^{|V|\times d}$ denote the node representation matrix at epoch t, and let $\mathcal{P}_k(\cdot)$ be a k-hop structural propagation operator. We define the structural resonance score of node v_i as

$$S_k^{(t)}(v_i) = \left\| \left[\mathcal{P}_k(H^{(t)}) - \mathcal{P}_k(H^{(t-1)}) \right]_i \right\|_2 . \tag{1}$$

Micro-level structural resonance holds at scale k if, for some epoch t, $\mathbb{E}_{v\sim P_{\text{ID}}} S_k^{(t)}(v) > \mathbb{E}_{v\sim P_{\text{OOD}}} S_k^{(t)}(v)$. Under this resonance pattern, smaller structural responses provide evidence for OOD candidate identification.

Instantiation via neighborhood propagation. We instantiate $\mathcal{P}_k(H) = A^k H$, where A is the row-normalized adjacency matrix with self-loops and is fixed across training epochs. Writing $\Delta H^{(t)} = H^{(t)} - H^{(t-1)}$, we obtain

$$S_k^{(t)}(v_i) = \left\| (A^k \Delta H^{(t)})_i \right\|_2 = \left\| \sum_j [A^k]_{ij} \Delta h_j^{(t)} \right\|_2 . \tag{2}$$

Here, $\Delta h_j^{(t)} = h_j^{(t)} - h_j^{(t-1)}$ is the optimization-induced representation perturbation of node v_j. The graph topology determines how these perturbation vectors are combined into a local resultant response.

Directional interactions beyond scalar score propagation. Following RSL, node-wise feature resonance is measured by $R^{(t)}(v_j) = \|\Delta h_j^{(t)}\|_2$. Propagating these pre-computed scalar scores would instead yield

$$Q_k^{(t)}(v_i) = \sum_j [A^k]_{ij} R^{(t)}(v_j) = \sum_j [A^k]_{ij} \|\Delta h_j^{(t)}\|_2 . \tag{3}$$

This operation discards perturbation directions before aggregation. In contrast, $S_k^{(t)}$ combines perturbation vectors before taking their norm, preserving their directional interactions.

To make this distinction explicit, let $a_{ij}^{(k)} = [A^k]_{ij}$. Expanding the squared structural resonance gives

$$\begin{aligned} \left(S_k^{(t)}(v_i)\right)^2 &= D_{i,k}^{(t)} + C_{i,k}^{(t)}, \qquad D_{i,k}^{(t)} = \sum_j \left(a_{ij}^{(k)}\right)^2 \left\| \Delta h_j^{(t)} \right\|_2^2 , \\ C_{i,k}^{(t)} &= 2 \sum_{j<\ell} a_{ij}^{(k)} a_{i\ell}^{(k)} \left\langle \Delta h_j^{(t)}, \Delta h_\ell^{(t)} \right\rangle . \end{aligned} \tag{4}$$

The term $D_{i,k}^{(t)}$ captures individual perturbation strengths, whereas $C_{i,k}^{(t)}$ captures pairwise directional interactions weighted by the propagation topology. Aligned perturbations reinforce one another, oppositely directed perturbations partially cancel, and orthogonal pairs contribute zero to the interaction term. Consequently, neighborhoods with identical perturbation magnitudes can have different structural resonance scores when their directional configurations differ. This information cannot be recovered from node-wise magnitudes alone.

The distinction therefore lies in the response being measured: MSRE uses the topology-conditioned interaction of inter-epoch representation perturbations as OOD evidence. The perturbations arise from consecutive training epochs, while their directional interactions are evaluated across nodes within each propagated neighborhood. Structural resonance does not require modifying the graph or assuming that OOD nodes exhibit anomalous raw topology statistics.

Implications for ID–OOD separability. The decomposition above explains what structural resonance retains; whether it improves ID–OOD discrimination depends on how these interactions differ between the two populations. When ID neighborhoods exhibit more coherent optimization responses, their perturbations can reinforce one another, whereas less aligned responses can undergo partial cancellation. Under such conditions, structural resonance can provide discriminative information beyond individual response magnitudes. Appendix B further discusses the role of neighborhood composition and retained self-information.

Directional cancellation also explains why structural resonance can weaken under noisy or heterophilic connectivity. By the reverse triangle inequality,

$$S_k^{(t)}(v_i) \geq c_{i,k} R^{(t)}(v_i) - \epsilon_{i,k}^{(t)}, \tag{5}$$

where $c_{i,k} = (A^k)_{ii}$ and $\epsilon_{i,k}^{(t)} = \sum_{j \neq i} (A^k)_{ij} \|\Delta h_j^{(t)}\|_2$. These are analytical quantities determined by the propagation matrix and perturbations, rather than tunable hyperparameters. The bound is informative when the retained self-response exceeds the neighborhood contribution; self-loops alone do not guarantee ID–OOD separability. This motivates using shallow propagation and combining structural resonance with node-wise feature resonance.

Joint resonance scoring. We integrate individual perturbation magnitudes and topology-conditioned resultant responses into

$$J^{(t)}(v_i) = \beta R^{(t)}(v_i) + (1-\beta)\sum_{k=1}^{K} S_k^{(t)}(v_i), \tag{6}$$

where $\beta \in [0,1]$ balances the two signals and K is the maximum propagation depth. The corresponding ID–OOD expectation gaps satisfy $G_J^{(t)} = \beta G_R^{(t)} + (1-\beta)\sum_{k=1}^{K} G_{S,k}^{(t)}$, as detailed in Appendix B. The feature term retains each node's response magnitude, while the structural terms additionally capture the directional reinforcement and cancellation of neighboring perturbations. Since these responses can fluctuate during optimization, we next aggregate the joint resonance across training epochs.

3.2 Multi-Scale Temporal Aggregation

Although joint resonance $J^{(t)}(v_i)$ provides a per-epoch measure of optimization-induced sensitivity, it is computed from local representation changes and may be affected by short-term fluctuations introduced by stochastic optimization. As a result, resonance values at individual epochs can be unstable and may not reliably reflect a node's distributional characteristics.

More importantly, OODness is not expected to manifest as a single isolated event during training. Instead, OOD nodes tend to exhibit persistently weak or inconsistent resonance responses over multiple consecutive epochs, whereas ID nodes maintain relatively stable resonance behavior under repeated optimization updates. Therefore, reliable OOD estimation requires aggregating resonance signals over a temporal neighborhood rather than relying on a single-epoch observation.

Temporal Resonance Sequence. Given a node v_i and a reference epoch t, we consider its recent temporal resonance sequence

$$\left[J^{(t-\tau)}(v_i),\ \ldots,\ J^{(t-1)}(v_i),\ J^{(t)}(v_i),\right], \tag{7}$$

which captures the evolution of joint resonance over a sliding temporal window. τ controls the size of the temporal window and determines the temporal scale for aggregating resonance behavior.

Resonance Aggregation. Based on the temporal resonance sequence, we define the aggregated temporal resonance score for node v_i at epoch t as

$$\widetilde{J}^{(t)}(v_i) = \sum_{s=0}^{\tau} J^{(t-s)}(v_i). \tag{8}$$

This sliding-window aggregation suppresses short-term fluctuations by integrating resonance signals over multiple consecutive epochs, while preserving persistent response patterns induced by optimization dynamics.

In principle, one could follow RSL Yang et al. (2025) to select a resonance-critical epoch using a labeled validation set, by scanning training epochs and choosing the stage that yields the best ID–OOD separation on validation nodes according to the resonance-based ranking. Then, OOD candidates can be directly obtained by taking the weakest-resonance nodes at the selected epoch, forming V_{cand}. However, this validation-dependent epoch selection is infeasible in realistic unlabeled deployments, where no reliable ID/OOD validation labels can be assumed. This motivates our validation-free selection strategy, which leverages temporal smoothing and cross-epoch consensus to identify persistently weak-resonance nodes without any validation supervision.

3.3 Validation-Free OOD Candidate Selection

Existing resonance-based OOD detectors often rely on a labeled validation set to select a resonance-critical epoch, which is impractical when ID/OOD validation labels are unavailable. Moreover, resonance measurements at a single epoch can be brittle, since GNN optimization is affected by

stochasticity and transient gradient noise, leading to unstable node rankings across epochs. To address these issues, we propose a validation-free candidate selection mechanism that identifies nodes with persistently weak resonance through temporal smoothing and cross-epoch consensus.

Let $J_i^{(t)}$ denote the joint resonance score of node v_i at epoch t, and let M be the candidate budget, i.e., the number of OOD candidates to select. Since weaker resonance indicates a higher likelihood of being OOD, candidate selection is based on persistently small resonance scores.

Temporal smoothing. We first apply exponential moving average (EMA) to stabilize epoch-wise resonance signals:

$$\widetilde{J}_i^{(t)} = \lambda J_i^{(t)} + (1-\lambda)\widetilde{J}_i^{(t-1)}, \tag{9}$$

where $\lambda \in (0,1)$ controls the smoothing strength. This suppresses transient fluctuations while preserving the overall weak-resonance trend of each node.

Cross-epoch consensus. To avoid selecting nodes that only accidentally receive low scores at a single epoch, we further measure the persistence of weak resonance across training. At each epoch t, nodes are ranked in ascending order according to $\widetilde{J}^{(t)}$, so that weaker resonance receives a smaller rank. Each node in the bottom-M weak-resonance set receives one vote, and its cumulative vote count is

$$V_i = \sum_{t=1}^{T} \mathbf{1}\{\mathrm{rank}_t(i) \le M\}, \tag{10}$$

where $\mathrm{rank}_t(i)$ denotes the rank of node v_i based on $\widetilde{J}_i^{(t)}$ at epoch t. A larger V_i indicates that node v_i is repeatedly identified as weak-resonance across epochs.

Dual-consensus selection. We prioritize nodes supported by both final EMA scores and cross-epoch votes. Let $L = \min(2M, |V|)$ and define $\mathcal{C}_{\mathrm{EMA}} = \mathrm{Bottom}_L(\widetilde{J}^{(T)})$ and $\mathcal{C}_{\mathrm{vote}} = \mathrm{Top}_M(V)$, containing the lowest-score and highest-vote nodes, respectively. Starting from their intersection $\mathcal{I} = \mathcal{C}_{\mathrm{EMA}} \cap \mathcal{C}_{\mathrm{vote}}$, we obtain

$$V_{\mathrm{cand}} = \mathcal{I} \cup \mathrm{Bottom}_{M-|\mathcal{I}|}\big(\widetilde{J}^{(T)}; \mathcal{C}_{\mathrm{EMA}} \setminus \mathcal{I}\big), \tag{11}$$

where the second term selects the remaining $M - |\mathcal{I}|$ nodes with the lowest final EMA scores from $\mathcal{C}_{\mathrm{EMA}} \setminus \mathcal{I}$. No additional nodes are selected when $|\mathcal{I}| = M$. Ties are broken by node indices. The expanded pool allows votes to influence selection, while filling the remaining slots preserves the budget of M without validation labels.

3.4 Final OOD Inference

Given the epoch-wise joint resonance scores $\{J^{(t)}\}_{t=1}^{T}$, we first obtain the OOD candidate set V_{cand} using the validation-free dual-consensus strategy in Sec. 3.3. We then use these candidates to generate proxy-OOD samples and train the final detector.

To translate the joint resonance signals into training supervision for the final detector, we use the selected candidates to guide Langevin-based proxy-OOD generation Yang et al. (2025), denoted by $\widetilde{X} = \mathrm{Langevin}(X, V_{\mathrm{cand}})$, yielding the proxy-OOD set $\widetilde{V}_{\mathrm{OOD}}$. The final OOD classifier $f(\cdot)$ is trained to distinguish known ID nodes from proxy-OOD nodes:

$$\mathcal{L}_{\mathrm{final}} = \mathbb{E}_{v \in V_{\mathrm{ID}}}\big[\ell(f(v), 0)\big] + \mathbb{E}_{\tilde{v} \in \widetilde{V}_{\mathrm{OOD}}}\big[\ell(f(\tilde{v}), 1)\big]. \tag{12}$$

At inference time, $f(v)$ is used as the OOD confidence score. The final prediction is obtained by thresholding: $\widehat{y}(v) = \mathrm{OOD}$ if $f(v) \ge \zeta$, and $\widehat{y}(v) = \mathrm{ID}$ otherwise, where ζ is the threshold used for FPR95 evaluation.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on a diverse collection of real-world graph datasets for node-level OOD detection, covering both multi-category citation and web graphs, as well as large-scale

Table 1: Main results of unsupervised node-level OOD detection on five real-world datasets. The best results are shown in bold, and the second-best results are underlined.

Method	Squirrel			WikiCS			YelpChi			Amazon			Reddit		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
LOF-KNN	51.85	29.87	95.21	44.06	37.48	96.28	56.27	27.86	93.91	55.91	32.40	88.62	53.73	5.15	93.42
MSP	53.63	30.98	94.59	57.96	48.43	92.11	58.31	29.11	92.47	68.80	45.73	85.70	57.96	6.31	92.98
ODIN	53.63	30.98	94.59	57.94	48.38	92.13	58.28	29.09	92.49	68.80	45.73	85.31	57.94	6.29	92.94
Energy	54.09	32.17	93.86	59.20	50.62	88.06	55.39	27.88	91.63	72.38	54.09	88.53	59.20	7.21	91.15
GKDE	50.14	29.92	92.93	65.48	69.92	80.42	-	-	-	-	-	-	-	-	-
OODGAT	59.67	31.02	94.43	84.33	62.39	88.27	58.28	29.11	94.43	88.27	58.28	84.33	59.67	7.11	94.43
GNNSafe	50.42	27.89	100.00	84.85	60.15	62.47	63.70	36.52	83.70	87.52	60.15	62.47	50.42	6.78	98.76
EnergyDef	**64.15**	**37.40**	**91.77**	70.22	60.10	83.17	62.04	29.71	90.62	86.57	74.50	32.43	63.32	8.34	89.34
RSL	62.28	36.06	93.85	84.01	81.14	49.23	65.59	36.77	83.70	90.03	83.91	19.60	64.83	10.18	89.84
MSRE (ours)	63.72	37.11	91.96	**88.24**	**84.40**	**47.29**	**67.26**	**37.47**	**82.43**	**96.48**	**90.04**	**13.97**	**65.97**	**10.82**	**86.69**

interaction and fraud detection networks. Specifically, we evaluate on two multi-category node classification benchmarks, including Squirrel and WikiCS Mernyei & Cangea (2020). These datasets exhibit varying degrees of homophily and heterophily, providing a comprehensive testbed for evaluating the robustness of unsupervised OOD detection methods under different structural regimes.

In addition, we include three binary classification datasets commonly used for fraud and anomaly detection: YelpChi Rayana & Akoglu (2015), Amazon McAuley & Leskovec (2013), and Reddit Kumar et al. (2019). These graphs are characterized by large scale, high connectivity, and severe class imbalance, posing significant challenges for OOD detection in realistic settings. Consistent with the unsupervised protocol, ground-truth labels are used only for evaluation and are never accessed during training.

Baselines. We compare MSRE with nine representative baselines covering local-density estimation, classifier-based scoring, uncertainty estimation, and graph-specific OOD detection. LOF-KNN Breunig et al. (2000) measures local-density deviations relative to nearest neighbors. MSP Hendrycks & Gimpel (2017) uses maximum softmax probability, ODIN Liang et al. (2018) combines temperature scaling with input perturbation, and Energy Liu et al. (2020) derives detection scores from classifier logits. GKDE Zhao et al. (2020) estimates node-level Dirichlet distributions using graph-based evidence from labeled nodes and supports uncertainty-based OOD detection. OODGAT Song & Wang (2022) uses attention to model interactions between ID and OOD nodes during message passing, while GNNSafe Wu et al. (2023) combines GNN-derived energy scores with energy belief propagation. We also include EnergyDef Gong & Sun (2024), a category-free energy-based detector, and RSL Yang et al. (2025), the closely related feature-resonance baseline.

Experimental Details. We follow the same dataset construction protocol and experimental settings as RSL. Detailed configurations are provided in Appendix C (*Experiment Details*).

Evaluation Metrics. Detection performance is evaluated using standard metrics, including the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision–Recall curve (AUPR), and the False Positive Rate at 95% True Positive Rate (FPR95). Higher AUROC and AUPR indicate better detection performance, while lower FPR95 is preferred. All reported results are averaged over multiple random splits.

4.2 Main Results

Table 1 compares MSRE with nine representative baselines under the experimental protocol adopted from RSL Yang et al. (2025). To align the resonance-stage selection setting with RSL, we report the validation-assisted version of MSRE here, which uses labeled validation data to select the resonance-critical training stage. This comparison evaluates the proposed structural-resonance framework under a matched selection protocol. The validation-free version is evaluated separately in Table 4 to assess whether competitive performance can be maintained without validation-based selection. Under the setting reported here, MSRE achieves the best AUROC, AUPR, and FPR95 on WikiCS, YelpChi, Amazon, and Reddit, and ranks second on all three metrics on Squirrel.

The comparison with RSL is particularly informative because it directly assesses the gains over the closely related feature-resonance baseline. MSRE improves upon RSL across all five datasets and all three metrics. The largest gains occur on Amazon, where AUROC increases from 90.03% to 96.48%, AUPR from 83.91% to 90.04%, and FPR95 decreases from 19.60% to 13.97%. On WikiCS, MSRE improves AUROC by 4.23 percentage points and AUPR by 3.26 percentage points,

while reducing FPR95 by 1.94 percentage points. These gains extend to YelpChi and Reddit, where all three metrics improve, including a 3.15-percentage-point reduction in FPR95 on Reddit. The improvements therefore span multiple datasets and capture both better overall ranking and lower false-positive rates at 95% true-positive rate.

Compared with other baselines, MSRE consistently outperforms MSP, ODIN, and Energy across all datasets and metrics, and outperforms graph-specific approaches on four benchmarks. Squirrel is an exception: although MSRE improves upon RSL in all three metrics, it remains slightly behind EnergyDef. The gaps are only 0.43, 0.29, and 0.19 percentage points in AUROC, AUPR, and FPR95, respectively. Thus, MSRE combines consistent gains over RSL with competitive performance against the broader set of baselines, while its relative advantage varies across datasets.

The following ablation studies isolate the contributions of structural directional interactions, temporal aggregation, and subsequent detection stages. They also compare validation-assisted and validation-free candidate selection, clarifying the performance trade-off when labeled validation data are unavailable. The additional computational cost remains limited on the evaluated benchmarks, with complexity analysis and empirical runtime and memory comparisons provided in Appendix F.

4.3 Ablation Studies

Table 2: Ablation study on structural resonance and directional interactions. The two variants remove structural resonance or retain only perturbation magnitudes, respectively.

Method	Squirrel			WikiCS			YelpChi			Amazon			Reddit		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
MSRE w/o structural	62.41	36.32	93.14	84.37	81.26	49.87	65.88	36.84	83.11	90.21	84.05	19.14	64.92	10.26	89.27
MSRE w/o directional interactions	63.26	36.64	92.61	86.51	82.05	48.53	65.92	36.86	83.52	91.95	85.93	17.14	65.06	10.59	88.51
MSRE (ours)	**63.72**	**37.11**	**91.96**	**88.24**	**84.40**	**47.29**	**67.26**	**37.47**	**82.43**	**96.48**	**90.04**	**13.97**	**65.97**	**10.82**	**86.69**

Effect of structural resonance and directional interactions. Table 2 examines the contribution of structural resonance and the directional interactions it retains. We compare full MSRE with two variants: one removes structural resonance entirely, while the other replaces $S_k^{(t)}(v_i) = \sqrt{D_{i,k}^{(t)} + C_{i,k}^{(t)}}$ with $\sqrt{D_{i,k}^{(t)}}$, retaining individual perturbation strengths but removing their pairwise directional interactions defined in Sec. 3.1. Full MSRE outperforms both variants across all five datasets in AUROC, AUPR, and FPR95. Removing structural resonance reduces AUROC from 96.48% to 90.21% on Amazon and from 88.24% to 84.37% on WikiCS, supporting the complementary value of topology-conditioned responses beyond node-wise perturbation magnitudes. Moreover, retaining only the magnitude term does not recover the performance of full MSRE. On Amazon, removing directional interactions reduces AUROC from 96.48% to 91.95% and increases FPR95 from 13.97% to 17.14%. On WikiCS, AUROC decreases from 88.24% to 86.51% and AUPR from 84.40% to 82.05%. Consistent degradations are also observed on the other datasets. These results support the mechanism described in Sec. 3.1: topology-conditioned reinforcement and cancellation among inter-epoch representation perturbations provide useful OOD evidence beyond the aggregation of perturbation strengths alone.

Table 3: Ablation study on multi-stage temporal resonance. Removing multi-stage temporal aggregation corresponds to using only single-stage resonance scores.

Method	Amazon			Reddit			WikiCS			Squirrel			YelpChi		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
MSRE w/o temporal	96.14	89.68	14.59	65.39	9.52	89.51	88.18	84.26	47.61	63.27	37.02	92.07	67.02	37.43	83.75
MSRE (ours)	**96.48**	**90.04**	**13.97**	**65.97**	**10.82**	**86.69**	**88.24**	**84.40**	**47.29**	**63.72**	**37.11**	**91.96**	**67.26**	**37.47**	**82.43**

Effect of temporal resonance aggregation. Removing multi-stage temporal resonance aggregation consistently degrades performance across all evaluated datasets. Compared with the full MSRE, the single-stage variant exhibits clear reductions in AUROC on Amazon (96.48%→96.14%), Reddit (65.97%→65.39%), and WikiCS (88.24%→88.18%), accompanied by noticeable decreases in AUPR and increases in FPR@95. The degradation is particularly pronounced on Reddit and YelpChi, where stochastic optimization noise and class imbalance are more severe. These results demonstrate that aggregating resonance signals across multiple temporal stages effectively suppresses transient fluctuations and yields more stable OOD scores.

Table 4: Ablation study on validation-free resonance stage selection. The variant removes reliance on validation-based epoch selection.

Method	Amazon			Reddit			WikiCS			Squirrel			YelpChi		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
RSL	90.03	83.91	19.60	64.83	10.18	89.84	84.01	81.14	49.23	62.28	36.06	93.85	65.59	36.77	83.70
MSRE w/ val-set	96.48	90.04	13.97	65.97	10.82	86.69	88.24	84.40	47.29	63.72	37.11	91.96	67.26	37.47	82.43
MSRE w/o val-set	95.43	87.47	18.57	65.28	10.65	88.31	87.08	83.01	48.42	63.13	36.78	92.04	66.67	36.72	83.81

Effect of validation-free candidate selection. Table 4 evaluates the impact of removing validation-based resonance stage selection. Without relying on a validation set, MSRE maintains competitive performance across all datasets, with only moderate decreases in AUROC (e.g., -1.05% on Amazon and -1.16% on WikiCS) and limited changes in AUPR. This confirms that the proposed validation-free aggregation strategy provides a robust alternative to heuristic epoch selection while preserving most of the detection performance. In contrast to RSL, which depends on validation heuristics to identify a resonance-critical training stage, MSRE achieves stable performance under unsupervised settings, making it more suitable for real-world deployment.

Table 5: Ablation study on proxy-OOD generation and final inference. Direct resonance scoring uses the aggregated resonance scores for OOD inference without proxy-OOD generation or classifier training. The variant without proxy-OOD generation trains the classifier using the selected candidates directly.

Method	Squirrel			WikiCS			YelpChi			Amazon			Reddit		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
Direct resonance scoring	60.63	36.20	93.16	82.51	80.05	69.53	65.17	36.02	85.91	92.78	89.97	14.15	52.85	7.11	92.77
MSRE w/o proxy-OOD	59.74	35.03	93.68	83.41	81.50	50.86	65.43	36.49	83.37	93.01	89.91	23.75	62.22	8.47	90.52
MSRE (ours)	**63.72**	**37.11**	**91.96**	**88.24**	**84.40**	**47.29**	**67.26**	**37.47**	**82.43**	**96.48**	**90.04**	**13.97**	**65.97**	**10.82**	**86.69**

Effect of proxy-OOD generation and final inference. Table 5 examines the detection stages following resonance estimation. We compare full MSRE with direct resonance scoring, which uses aggregated resonance scores for OOD inference, and MSRE w/o proxy-OOD, which trains the classifier using selected candidates directly alongside known ID nodes. Full MSRE outperforms both variants across all five datasets in AUROC, AUPR, and FPR95. Compared with direct scoring, it improves AUROC from 82.51% to 88.24% on WikiCS and from 52.85% to 65.97% on Reddit, while reducing FPR95 on WikiCS from 69.53% to 47.29%. These results support using resonance for candidate discovery, with subsequent proxy-OOD generation and classifier training providing additional gains beyond the resonance ranking.

Removing proxy-OOD generation while retaining the classifier also consistently degrades performance. On Amazon, AUROC decreases from 96.48% to 93.01%, and FPR95 increases from 13.97% to 23.75%. On WikiCS, AUROC drops from 88.24% to 83.41%. Moreover, the classifier trained directly on selected candidates does not uniformly outperform direct scoring, as illustrated by its higher FPR95 on Amazon (23.75% versus 14.15%). The gains therefore cannot be attributed simply to adding a classifier. Instead, the results suggest that proxy-OOD generation supplements the limited training evidence provided by the selected candidates, supporting the complementary roles of resonance-based candidate discovery and subsequent training for effective ID–OOD discrimination.

Additional analyses are provided in the appendix. Appendix D examines sensitivity to the resonance balancing coefficient β and temporal window parameter τ, while Appendix E evaluates the effects of ID label availability and graph homophily.

5 Conclusion

We presented **Multi-Scale Structural Resonance Estimation (MSRE)**, an unsupervised framework for node-level out-of-distribution (OOD) detection on graphs. MSRE complements feature resonance with structural resonance by modeling local structure-propagated perturbation responses, and unifies them into a joint resonance score. To improve stability without validation supervision, we further introduce multi-scale temporal aggregation and a cross-epoch consensus strategy for reliable OOD candidate selection. Extensive experiments on multiple benchmark graphs demonstrate consistent improvements over state-of-the-art unsupervised and resonance-based baselines.

AI USE STATEMENT

We used generative AI tools to assist with LaTeX typesetting, language polishing and translation, and manuscript revision. These revisions included checking numerical consistency and correcting errors in wording and presentation. AI tools also assisted with clarifying mathematical explanations and interpreting reported experimental results. The authors take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This work studies node-level out-of-distribution detection using existing graph benchmarks, without recruiting human participants or collecting new personal data. We acknowledge the original data sources and prior methods. Any reuse or redistribution of benchmark data should respect the applicable licenses and access conditions, with appropriate precautions against re-identification or disclosure of potentially sensitive information.

This work focuses on methodological development and evaluation on existing graph benchmarks. It does not involve deployment in systems that make consequential decisions about individuals, nor does it attempt to recover personal identities or infer sensitive personal attributes. Within the scope of the research presented, we identify no additional fairness or privacy concerns introduced by the proposed method or experimental procedures.

To support transparent evaluation, we distinguish validation-assisted comparisons from validation-free evaluations and report computational overhead and limitations. The authors are responsible for accurately describing the experimental settings and results, acknowledging prior work, and disclosing AI assistance.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main text specifies the resonance scores, temporal aggregation, candidate selection, and final classifier training procedure. Appendix C describes the dataset splits, implementation settings, and shared hyperparameters. Additional derivations and assumptions for structural resonance are presented in Appendix B. Parameter sensitivity and computational overhead are examined in Appendices D and F, respectively, providing further information on the method's configuration and resource requirements.

REFERENCES

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104. Association for Computing Machinery, 2000. doi: 10.1145/342009.335388.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602. Society for Industrial and Applied Mathematics, 2019. doi: 10.1137/1.9781611975673.67.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

Zheng Gong and Ying Sun. An energy-centric framework for category-free out-of-distribution node detection in graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 908–919. Association for Computing Machinery, 2024. doi: 10.1145/3637528.3671939.

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.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

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

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1269–1278. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330895.

Zenan Li, Qitian Wu, Fan Nie, and Junchi Yan. GraphDE: A generative framework for debiased learning and out-of-distribution detection on graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30277–30290. Curran Associates, Inc., 2022. doi: 10.52202/068431-2195.

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.

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, pp. 21464–21475. Curran Associates, Inc., 2020.

Longfei Ma, Yiyou Sun, Kaize Ding, Zemin Liu, and Fei Wu. Revisiting score propagation in graph out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4341–4373. Curran Associates, Inc., 2024. doi: 10.52202/079017-0142.

Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 897–908, 2013. doi: 10.1145/2488388.2488466.

Péter Mernyei and Cătălina Cangea. Wiki-CS: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 985–994. Association for Computing Machinery, 2015. doi: 10.1145/2783258.2783370.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yu Song and Donglin Wang. Learning on graphs with out-of-distribution nodes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1635–1645. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539457.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11839–11852. Curran Associates, Inc., 2020.

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

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23631–23644. PMLR, 2022.

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.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56216–56234. PMLR, 2024.

Shenzhi Yang, Junbo Zhao, Sharon Li, Shouqing Yang, Dingyu Yang, Xiaofang Zhang, and Haobo Wang. Harnessing feature resonance under arbitrary target alignment for out-of-distribution node detection. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-5417.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12827–12836. Curran Associates, Inc., 2020.

Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T. Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12220–12233, 2023. doi: 10.1109/TKDE.2021.3119326.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. doi: 10.1016/j.aiopen.2021.01.001.

A Related Work

A.1 Out-of-Distribution Detection

OOD detection identifies inputs outside the target ID distribution, while open-set recognition additionally requires classification among known categories. Representative OOD approaches use classifier outputs, density estimates, or self-supervised representations. MSP Hendrycks & Gimpel (2017) uses maximum softmax probability, while ODIN Liang et al. (2018) improves score separation through temperature scaling and input perturbation. Energy-based detection Liu et al. (2020) constructs scores from classifier logits and can also incorporate energy-based training objectives. In contrast to post-hoc scoring, LogitNorm Wei et al. (2022) modifies classifier training by normalizing logits within the cross-entropy objective to mitigate overconfidence.

Density-based detection uses probabilistic models to assess whether samples are consistent with the ID distribution. However, deep generative models can assign higher likelihoods to OOD inputs than to ID inputs Nalisnick et al. (2019). To address confounding background statistics, Ren et al. Ren et al. (2019) propose likelihood-ratio scoring rather than relying on raw likelihood alone. Outlier Exposure Hendrycks et al. (2019) instead uses auxiliary outlier data to improve detection, although obtaining suitable external outliers can be costly.

Other approaches reduce dependence on external outliers. VOS Du et al. (2022) synthesizes virtual outliers from low-likelihood regions of estimated class-conditional feature distributions and uses them to regularize training. CSI Tack et al. (2020) learns representations through contrastive learning with distribution-shifting transformations and supports settings without category labels. These methods differ in their supervision requirements, but do not directly address the relational dependencies introduced by graph message passing. This motivates graph-specific detection mechanisms, including category-label-free signals derived from optimization-induced representation changes.

A.2 Graph Out-of-Distribution Detection

Graph OOD detection must account for dependencies introduced by graph connectivity and message passing. OODGAT Song & Wang (2022) uses attention to distinguish ID–OOD interactions during feature propagation. GNNSafe Wu et al. (2023) derives energy scores from GNN classifier outputs and improves detection through energy belief propagation. NodeSafe Yang et al. (2024) addresses extreme scores in energy aggregation by introducing regularization that bounds negative energy scores and mitigates logit shifts. GRASP Ma et al. (2024) studies when OOD score propagation is beneficial and introduces an edge augmentation strategy to improve propagation; its framework can be combined with different base scoring functions. These methods should not be characterized as uniformly requiring auxiliary OOD training data. Their use of ID category supervision or pretrained classifier-based scores nevertheless distinguishes their standard settings from the category-label-free setting considered here.

A.3 Unsupervised Graph Out-of-Distribution Detection

Recent progress has explored unsupervised OOD detection on graphs. Reconstruction-based paradigms predict abnormality based on attribute–structure reconstruction errors, as in DOMINANT Ding et al. (2019), while contrastive frameworks such as SL-GAD Zheng et al. (2023) learn discriminative structure-attribute embeddings without explicit labels. Energy-based methods, e.g., EnergyDef Gong & Sun (2024), learn graph energy surfaces to distinguish ID and OOD samples. Despite reducing label dependency, none of these approaches exploit gradient-induced resonance signals during optimization. RSL Yang et al. (2025) initiates this direction by measuring feature resonance through forced alignment to a target vector and identifying OOD nodes from oscillation magnitudes. However, prior resonance-based methods overlook structural resonance induced by message passing, rely on validation sets to select resonance phases, and evaluate resonance signals only at isolated training stages, ignoring their temporal evolution. These limitations motivate us to develop a new unsupervised graph OOD detection method that jointly models feature- and structure-level resonance signals and stabilizes resonance estimation across training stages without relying on validation supervision.

A.4 Feature Resonance

RSL Yang et al. (2025) identifies a label-agnostic phenomenon, termed *feature resonance*, for unsupervised graph OOD detection. The key observation is that when the GNN is optimized using objectives defined on a small set of known ID nodes, unknown ID nodes typically undergo larger representation updates than OOD nodes. Consequently, weaker representation dynamics are indicative of a higher likelihood of being OOD.

Definition 1 (Micro-level Feature Resonance Yang et al. (2025))**.** *Consider a GNN trained by an arbitrary objective defined on known ID nodes. Micro-level feature resonance holds if there exists an epoch t such that*

$$\mathbb{E}_{v\sim P_{\mathrm{ID}}}\left[\left\|h^{(t)}(v)-h^{(t-1)}(v)\right\|_2\right] > \mathbb{E}_{v\sim P_{\mathrm{OOD}}}\left[\left\|h^{(t)}(v)-h^{(t-1)}(v)\right\|_2\right]. \tag{13}$$

where $h^{(t)}(v)$ denotes the representation of node v at epoch t, and P_{ID} and P_{OOD} denote the underlying distributions of in-distribution and out-of-distribution nodes, respectively. Accordingly, the resonance score of node v_i at epoch t is defined as

$$R^{(t)}(v_i) = \left\|h^{(t)}(v_i)-h^{(t-1)}(v_i)\right\|_2, \tag{14}$$

where a smaller $R^{(t)}(v_i)$ indicates weaker resonance and thus a higher likelihood of being OOD.

To induce optimization-driven perturbations without category supervision, RSL aligns the representations of a small set of known ID nodes to an arbitrary target in the embedding space, which is instantiated in practice as an Equiangular Tight Frame (ETF). ETF targets provide uniformly distributed directions and act as a label-agnostic forcing signal that facilitates stable representation alignment. Based on the resulting resonance scores, RSL identifies nodes with weak resonance as candidate OOD samples, which are subsequently used to train a self-supervised ID/OOD classifier.

B Detailed Derivations for Structural Resonance

Relation to feature resonance. Let $\Delta H^{(t)} = H^{(t)} - H^{(t-1)}$ denote the optimization-induced representation perturbation matrix at epoch t. RSL measures node-wise feature resonance by $R^{(t)}(v_i) = \|\Delta h_i^{(t)}\|_2$. Our structural resonance generalizes this node-wise measurement to topology-aware perturbations:

$$S_k^{(t)}(v_i) = \|(A^k\Delta H^{(t)})_i\|_2 = \left\|\sum_j [A^k]_{ij}\Delta h_j^{(t)}\right\|_2. \tag{15}$$

Thus, structural resonance can be interpreted as a graph-propagated form of feature resonance, where the propagation operator aggregates perturbations over the k-hop neighborhood.

Mean-field approximation. To explain why structural resonance can enlarge ID–OOD separability, we consider a mean-field model during the resonance period. Assume that node-wise perturbations can be approximated as

$$\Delta h_i^{(t)} \approx \begin{cases} \mu_{\mathrm{ID}} + \xi_i^{(t)}, & v_i \in V_{\mathrm{ID}}, \\ \mu_{\mathrm{OOD}} + \zeta_i^{(t)}, & v_i \in V_{\mathrm{OOD}}, \end{cases} \tag{16}$$

where $\|\mu_{\mathrm{ID}}\|_2 > \|\mu_{\mathrm{OOD}}\|_2$, and $\xi_i^{(t)}$ and $\zeta_i^{(t)}$ are zero-mean residual terms. Let $\rho_{i,k} = \sum_{j\in V_{\mathrm{ID}}}[A^k]_{ij}$ denote the amount of ID-like resonance mass in the k-hop propagated neighborhood of v_i. Taking expectation gives

$$\mathbb{E}\Big[(A^k\Delta H^{(t)})_i\Big] = \rho_{i,k}\mu_{\mathrm{ID}} + (1-\rho_{i,k})\mu_{\mathrm{OOD}}. \tag{17}$$

Therefore, if ID nodes have larger local ID coherence than OOD nodes, i.e., $\rho_k^{\mathrm{ID}} > \rho_k^{\mathrm{OOD}}$, structural propagation increases the expected ID–OOD resonance gap. This explains why structural resonance can amplify separability when local topology preserves distribution-relevant neighborhood consistency.

Inner-product alignment decomposition. The above mean-field explanation is strongest under locally coherent neighborhoods. To analyze noisy or heterophilic cases, we decompose the propagated perturbation into a retained self-term and a neighborhood-induced term:

$$(A^k \Delta H^{(t)})_i = \alpha^k \Delta h_i^{(t)} + \eta_{i,k}^{(t)}. \tag{18}$$

Then

$$S_k^{(t)}(v_i)^2 = \alpha^{2k} \|\Delta h_i^{(t)}\|_2^2 + \|\eta_{i,k}^{(t)}\|_2^2 + 2\alpha^k \langle \Delta h_i^{(t)}, \eta_{i,k}^{(t)} \rangle. \tag{19}$$

Defining the alignment coefficient

$$\kappa_{i,k} = \frac{\langle \Delta h_i^{(t)}, \eta_{i,k}^{(t)} \rangle}{\|\Delta h_i^{(t)}\|_2 \|\eta_{i,k}^{(t)}\|_2}, \tag{20}$$

we obtain

$$S_k^{(t)}(v_i)^2 = \alpha^{2k} \|\Delta h_i^{(t)}\|_2^2 + \|\eta_{i,k}^{(t)}\|_2^2 + 2\alpha^k \|\Delta h_i^{(t)}\|_2 \|\eta_{i,k}^{(t)}\|_2 \kappa_{i,k}. \tag{21}$$

When $\kappa_{i,k} > 0$, the self-term and neighborhood term reinforce each other; when $\kappa_{i,k} \approx 0$, the structural signal is mainly preserved through the self-term; and when $\kappa_{i,k} < 0$, the neighborhood term partially cancels the self-term. Thus, heterophily may weaken structural resonance, but it does not necessarily eliminate it as long as the retained self-term remains non-negligible.

Lower bound under self-retention. Since A contains self-loops, A^k has a positive diagonal entry, i.e., $(A^k)_{ii} > 0$. Therefore,

$$S_k^{(t)}(v_i) = \left\| (A^k)_{ii} \Delta h_i^{(t)} + \sum_{j \neq i} (A^k)_{ij} \Delta h_j^{(t)} \right\|_2. \tag{22}$$

By the triangle inequality,

$$S_k^{(t)}(v_i) \geq (A^k)_{ii} \|\Delta h_i^{(t)}\|_2 - \sum_{j \neq i} (A^k)_{ij} \|\Delta h_j^{(t)}\|_2. \tag{23}$$

Let $c_{i,k} = (A^k)_{ii}$ and $\epsilon_{i,k}^{(t)} = \sum_{j \neq i} (A^k)_{ij} \|\Delta h_j^{(t)}\|_2$. Then

$$S_k^{(t)}(v_i) \geq c_{i,k} R^{(t)}(v_i) - \epsilon_{i,k}^{(t)}. \tag{24}$$

This lower bound shows that structural resonance preserves a scaled version of node-wise feature resonance, even when neighborhood aggregation introduces noise.

Gap decomposition. Define the feature-resonance gap as

$$G_R^{(t)} = \mathbb{E}_{v \sim P_{\mathrm{ID}}} R^{(t)}(v) - \mathbb{E}_{v \sim P_{\mathrm{OOD}}} R^{(t)}(v), \tag{25}$$

and the structural-resonance gap at scale k as

$$G_{S,k}^{(t)} = \mathbb{E}_{v \sim P_{\mathrm{ID}}} S_k^{(t)}(v) - \mathbb{E}_{v \sim P_{\mathrm{OOD}}} S_k^{(t)}(v). \tag{26}$$

Using the lower bound above, we obtain the informal relation

$$G_{S,k}^{(t)} \underline{c}_k G_R^{(t)} - \left(\bar{\epsilon}_{\mathrm{ID},k}^{(t)} + \bar{\epsilon}_{\mathrm{OOD},k}^{(t)} \right), \tag{27}$$

where $\underline{c}_k = \min_i c_{i,k}$, and $\bar{\epsilon}_{\mathrm{ID},k}^{(t)}$ and $\bar{\epsilon}_{\mathrm{OOD},k}^{(t)}$ denote average residual terms over ID and OOD nodes. Hence, whenever the retained self-term and feature-resonance gap dominate neighborhood noise, the structural gap remains positive and can further increase ID–OOD separability.

Table 6: Additional hyperparameter settings shared across datasets.

Hyperparameter	Symbol	Value
Structural propagation depth	K	2
EMA coefficient	λ	0.7
Candidate budget	M	4

Joint-score gap. For the joint resonance score

$$J^{(t)}(v_i) = \beta R^{(t)}(v_i) + (1-\beta)\sum_{k=1}^{K} S_k^{(t)}(v_i), \tag{28}$$

the corresponding ID–OOD gap decomposes as

$$G_J^{(t)} = \beta G_R^{(t)} + (1-\beta)\sum_{k=1}^{K} G_{S,k}^{(t)}. \tag{29}$$

Therefore, when local structural coherence is strong, structural resonance enlarges the resonance gap in graph space. When local coherence is weak, the node-wise resonance term preserves a stable gap. This also explains why shallow multi-scale propagation is preferable: small k prevents the self-retention coefficient from decaying too quickly and reduces the risk of over-smoothing.

C Experiment Details

We follow the same dataset construction protocol as RSL and adopt GCN as the backbone encoder. Under this setting, the node attributes of unlabeled/unknown nodes are accessible, and thus leveraging their features during training to identify reliable OOD candidates is a legitimate and realistic strategy. For structural resonance, we aggregate perturbation responses up to the 2-hop neighborhood. For temporal aggregation, we set the sliding-window size to $\tau = 2$. All experiments are conducted on a single NVIDIA RTX 4080 GPU.

Regarding dataset splits, for Squirrel and WikiCS, we randomly select one and two classes as OOD classes, respectively. For the fraud detection datasets, we treat legitimate entities as ID nodes and fraudsters as OOD nodes. Following RSL, we use 40% of ID nodes for training, and split the remaining nodes into validation and test sets with a 1:2 ratio via stratified sampling over ID/OOD labels.

We use a structural propagation depth of $K = 2$, an EMA coefficient of $\lambda = 0.7$, and a candidate budget of $M = 4$. These values are fixed across datasets rather than tuned separately using OOD validation labels. Shallow propagation captures local response interactions while limiting the propagation depth over which unrelated responses may be mixed. Under the EMA convention in Eq. 9, $\lambda = 0.7$ assigns a weight of 0.7 to the current resonance score and 0.3 to the previous smoothed score. The candidate budget controls the number of seeds used for proxy-OOD generation; it neither specifies the true OOD proportion nor determines the number of final OOD predictions.

D Sensitivity Analysis

Sensitivity of resonance balancing coefficient β. Table 7 analyzes the sensitivity of MSRE to the resonance balancing coefficient β. Over-emphasizing structural resonance (e.g., $\beta = 0.3$) leads to a substantial performance drop (AUROC: 96.48%→86.73%, FPR@95: 13.97%→26.91%), indicating that structural resonance alone is insufficient for reliable OOD detection. Conversely, increasing β toward feature-dominated settings (e.g., $\beta = 0.9$) also degrades performance (AUROC: 96.48%→90.17%, FPR@95: 13.97%→21.06%). These results highlight the necessity of jointly modeling feature- and structure-induced resonance for robust OOD detection on graph-structured data.

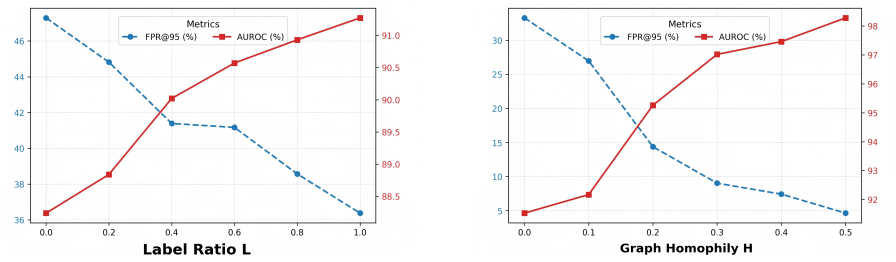
Sensitivity of sliding-window size τ. Table 7 evaluates the effect of the temporal aggregation window τ. A moderate window achieves the best performance: $\tau = 2$ yields the highest AUROC (96.48%) and lowest FPR@95 (13.97%). Smaller ($\tau = 1$) or larger windows lead to degraded results, suggesting that overly short aggregation misses temporal consistency while overly long aggregation over-smooths resonance dynamics. We set $\tau = 2$ in all experiments.

Table 7: Sensitivity analysis of key hyperparameters on Amazon. β balances feature and structural resonance, and τ controls the sliding-window size for temporal aggregation.

β	AUROC↑	FPR@95↓	τ	AUROC↑	FPR@95↓
0.3	86.73	26.91	1	96.14	14.59
0.5	**96.48**	**13.97**	2	**96.48**	**13.97**
0.7	94.55	17.04	3	94.29	17.02
0.9	90.17	21.06	4	90.17	20.58

E Generalization Analysis

Following the experimental protocol of RSL, we study how label availability and graph homophily influence the performance of our proposed method. Specifically, we evaluate the effect of the available ID label ratio L and the homophily level H, with results reported in Fig. 2a and Fig. 2b.



(a) Effect of L (b) Effect of H

Figure 2: (a) Effect of the available label ratio L on WikiCS; (b) Effect of graph homophily H on CiteSeer.

Effect of label availability. On WikiCS, increasing the proportion of available ID labels consistently improves OOD detection performance: As L increases from 0 to 1.0, AUROC improves from 88.24% to 91.27%. AUPR increases from 84.40% to 89.75%. FPR@95 decreases from 47.29% to 36.38%. This trend suggests that stronger supervision leads to better initialization of ID representations, which produces more coherent resonance responses during optimization and amplifies the separability between ID and OOD nodes. Notably, even in the unsupervised setting ($L = 0$), our method maintains strong performance, demonstrating that it does not rely on label supervision.

Effect of graph homophily. A similar monotonic improvement is observed on CiteSeer as graph homophily increases: AUROC improves substantially from 91.53% to 98.28%. FPR@95 is sharply reduced from 33.27% to 4.64%. Higher homophily strengthens neighborhood consistency and stabilizes perturbation propagation, thereby enhancing both feature- and structure-induced resonance signals. These results indicate that our method benefits from improved structural coherence, while remaining robust across a wide range of homophily levels. We further evaluate sensitivity to incomplete graph connectivity through random edge removal. The results show gradual performance degradation as the removal ratio increases, with details provided in Appendix G.

F Computational Complexity and Practical Overhead

Computational complexity. Let $N = |V|$, $m = |E|$, and d denote the number of nodes, the number of edges, and the representation dimension, respectively. Computing node-wise feature resonance requires $O(Nd)$ time per epoch. Structural responses up to depth K can be evaluated recursively as $Z_0 = \Delta H^{(t)}$ and $Z_k = AZ_{k-1}$, avoiding explicit construction of A^k. With sparse adjacency multiplication, these propagations and their row-wise norms require $O(K(m+N)d)$ time per epoch, including self-loops. Scalar EMA updates require $O(N)$ time. Sliding-window sums can also be updated in $O(N)$ time using a rolling sum, while candidate selection based on a full

Table 8: End-to-end runtime comparison in seconds, including final classifier training.

Method	Squirrel	WikiCS	YelpChi	Amazon	Reddit
RSL	42.45	45.38	99.32	95.67	82.83
MSRE	44.71	49.63	114.09	103.47	98.28

Table 9: Peak GPU memory comparison in MB.

Method	Squirrel	WikiCS	YelpChi	Amazon	Reddit
RSL	1246.81	437.30	928.32	984.95	70.30
MSRE	1417.76	451.47	1008.85	1001.04	108.88

ranking requires $O(N \log N)$ time per epoch. Under this implementation, resonance computation and selection therefore require

$$O(Nd + K(m+N)d + N \log N) \tag{30}$$

time per epoch. By processing propagation scales sequentially, the representation buffers require $O(Nd)$ memory; storing the scalar temporal window, EMA state, and vote counts requires $O(N(\tau+1))$ memory under the window definition in Eq. 8. These costs exclude graph storage, backbone training, proxy-OOD generation, and final classifier training.

Practical overhead. Tables 8 and 9 report runtime and peak GPU memory under the same experimental setting, with a visual comparison in Fig. 3. MSRE increases runtime by 5.3%–18.7% relative to RSL across the five datasets. For example, runtime increases from 42.45 to 44.71 seconds on Squirrel and from 82.83 to 98.28 seconds on Reddit. The memory overhead varies across datasets. On Amazon, peak memory increases from 984.95 to 1001.04 MB, whereas on Reddit it increases from 70.30 to 108.88 MB. These measurements characterize the additional runtime and memory cost of the complete MSRE pipeline on the evaluated benchmarks.

G Robustness to Random Edge Removal

Effect of random edge removal. Table 10 evaluates MSRE after randomly removing 10%, 20%, and 30% of graph edges. AUROC decreases progressively with increasing removal ratios across all five datasets. At 20% removal, AUROC decreases from 88.24% to 84.93% on WikiCS and from 96.48% to 92.73% on Amazon. Further removal leads to larger declines, highlighting the dependence of structural responses on the available graph connectivity. These results complement the clean-graph evaluation by showing gradual performance degradation under the tested random edge-removal conditions.

H Limitations

MSRE relies on structure-induced resonance signals obtained from shallow neighborhood propagation. Thus, the informativeness of structural resonance may vary with graph connectivity quality and neighborhood consistency. In practice, MSRE combines structural resonance with node-wise feature resonance, which helps maintain stable behavior when graph structures are noisy or less homophilic.

MSRE also tracks resonance signals across training epochs and performs temporal aggregation. Although this improves robustness to transient optimization noise, it introduces additional bookkeeping compared with single-epoch scoring methods. We keep this overhead moderate by using shallow propagation and a small temporal window, and future work may explore more efficient resonance caching or adaptive temporal sampling for larger graphs.

Finally, our experiments follow established node-level graph OOD detection protocols on widely used benchmark datasets. Further evaluation on larger, dynamic, or heterogeneous graphs would be valuable for assessing the broader applicability of resonance-based OOD detection.

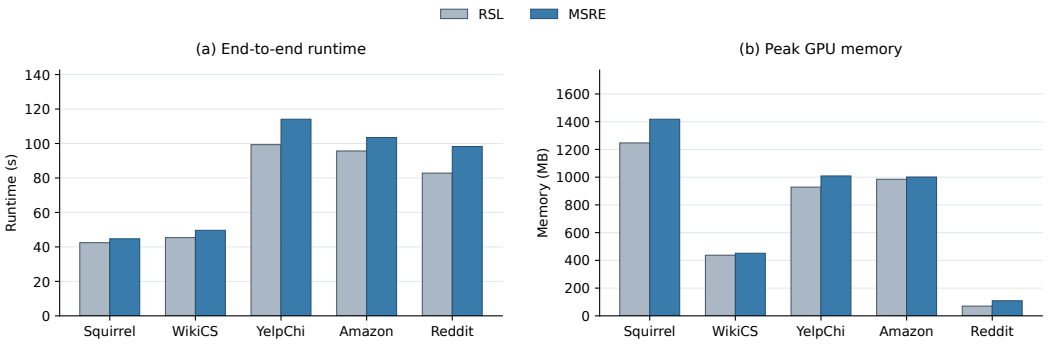


Figure 3: Computational overhead of RSL and MSRE across five datasets. (a) End-to-end runtime. (b) Peak GPU memory. Exact measurements are provided in Tables 8 and 9.

Table 10: AUROC (%) under different random edge-removal ratios. The 0% setting uses the original graph.

Dataset	0%	10%	20%	30%
Squirrel	63.72	62.07	61.42	59.64
WikiCS	88.24	86.56	84.93	81.28
YelpChi	67.26	66.17	64.45	61.21
Amazon	96.48	95.33	92.73	90.16
Reddit	65.97	64.84	63.02	60.79

I Broader Impacts

This work aims to improve the reliability of graph neural networks under label-scarce graph OOD detection scenarios. By identifying nodes whose feature or structural behavior deviates from the in-distribution population, MSRE may help reduce unreliable predictions in applications such as fraud detection, recommender systems, cybersecurity monitoring, and molecular property analysis. Since the proposed method does not require semantic category labels or labeled validation data, it may also be useful in practical settings where annotation is costly or unavailable.

At the same time, OOD detection methods should be deployed with appropriate caution. An improved detector may help flag potentially abnormal or unreliable instances, but it should not be treated as a standalone decision-making system in high-stakes domains. False positives may incorrectly mark normal nodes as suspicious, while false negatives may fail to identify truly anomalous nodes. Therefore, in sensitive applications, the detector should be used as a supportive reliability module together with domain expertise, auditing procedures, and task-specific safeguards. The proposed method is evaluated on benchmark datasets and does not introduce new data collection, human-subject experiments, or high-risk generative capabilities.