# Semantically Complementary Spectral Views Learning for Graph-Level Anomaly Detection

**Yuanming Liu**[1*], **Fan Xu**[5*], **Xibin Zhao**[4], **Hai Wan**[4], **Junran Wu**[6], **Nan Wang**[2,3†], **Jiqiang Liu**[3]
[1]School of Software, Beijing Jiaotong University
[2]State Key Laboratory of Internet Architecture, Tsinghua University
[3]School of Cyberspace Science and Technology, Beijing Jiaotong University
[4]BNRist, KLISS, the School of Software, Tsinghua University
[5]AI4SE, Shenzhen Loop Area Institute
[6]School of Computing, National University of Singapore
[*] Co-first authors. [†] Corresponding author.
Main Contact: liuyuanming@bjtu.edu.cn

## Abstract

Graph-level anomaly detection (GLAD) is a critical task in domains like social networks and bioinformatics. However, current unsupervised contrastive learning-based methods rely on semantically homogeneous augmentations, leading to catastrophic informational overlap that trivializes the learning objective and effectively masks anomalous signals in GLAD. To overcome this, we propose **Sica** (**S**eman**i**cally **C**omplementary Spectr**a**l Views), a novel contrastive framework that learns from semantically complementary spectral views rather than redundant homogeneous views. **Sica** first constructs a pair of spectral views to separately capture global community structure and local heterophilous variations, which together cover the full graph spectrum of anomalous patterns. Inspired by the Multi-View Information Bottleneck principle, we further design a soft subgraph extraction module with *Score Reconstruction Attention (SRA)* and an *anchor-guided condensation mechanism* to distill the essential shared information from each view into a compact representation. We then optimize a contrastive objective across these distilled views to enforce cross-view alignment and amplify subtle structural deviations. Extensive experiments on challenging GLAD and graph-level out-of-distribution detection (GLOD) benchmarks demonstrate that **Sica** achieves substantial and consistent gains in average performance over state-of-the-art baselines. The code is available at https://github.com/Lqyrmk/sica.

## 1 Introduction

Graphs have become a cornerstone for modeling complex relational systems, from social and financial networks to biological interaction pathways [5, 43]. The advent of Graph Neural Networks (GNNs) has provided a powerful paradigm for learning from such structured data, fueling progress across numerous applications [12, 7, 41, 40]. A critical task within this domain is graph-level anomaly detection (GLAD), which seeks to identify entire graphs that deviate from the normative pattern [22, 47, 38, 3]. The importance of GLAD is profound, spanning the detection of coordinated fraudulent activities in transaction networks, the identification of toxic molecules in bioinformatics [39], and the discovery of unusual events in transportation systems [25].

Despite its critical importance, GLAD remains challenging due to the scarcity of labeled anomalous data. In real-world scenarios, anomalies are rare and costly to annotate. Therefore, recent research has turned to unsupervised methods [48, 13], particularly Graph Contrastive Learning (GCL) [4, 8, 17, 13], which aims to learn invariant representations by maximizing consistency between augmented views. However, the efficacy of GCL hinges critically on the quality of views. Existing methods based on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

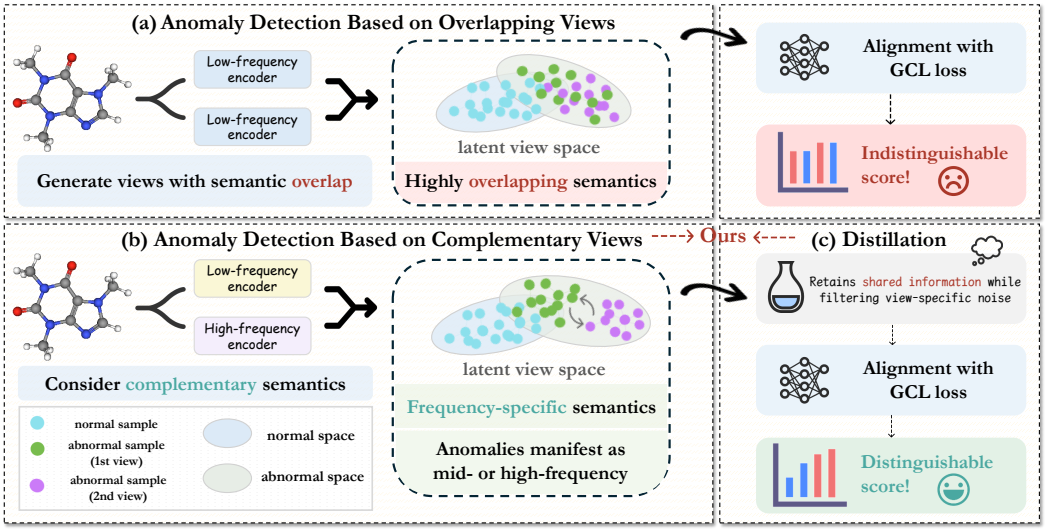



Figure 1: (a) Overlapping homogeneous semantics lead to consistent patterns, which tend to cause catastrophic shortcuts in contrastive learning and thus weaken discriminative ability. (b) Complementary spectral semantics can capture unique high-frequency anomaly patterns. (c) Distillation for views can eliminate semantic noise and extract core information shared between these views.

random perturbation [46] or deterministic transformation [19, 18, 9] often generate *semantically homogeneous views*. As shown in Figure 1(a), these views originate from similar spaces with low-frequency encoders, capturing *overlapping semantics* and focusing only on dominant low-frequency graph patterns. This semantic overlap presents a fatal flaw: the contrastive objective becomes trivially easy to optimize, leading to both normal and abnormal samples exhibiting high cross-views consistency. This blurs the decision boundary and weakens the discriminative ability of model.

To better leverage contrastive learning to distinguish anomalies, we believe it is necessary to move beyond semantically overlapping views and instead construct *semantically complementary views*. We draw inspiration from graph spectral theory [30], which suggests that essential properties of a graph are encoded in its spectral signature. Crucially, anomalies often manifest as aberrations in this signature. For instance, an abnormally cohesive fraudulent subgroup can be seen as a low-frequency anomaly, while unusual long-range connections bridging disparate communities represent a high-frequency anomaly. Traditional methods inherently tend to detect the former while ignoring the latter due to the low-pass nature. This motivates our core idea in Figure 1(b): to effectively identify the full spectrum of anomalies, we must analyze the graph from both low-frequency and high-frequency perspectives simultaneously. However, we notice that the difficulty of this idea lies in that each complementary view contains *view-specific noise*, resulting in limited shared information among views, which makes direct contrastive learning challenging [4]. Therefore, as shown in Figure 1(c), we need a special approach to act as a "distillation module" for views, to eliminate view-specific noise and extract the essential information shared between these complementary views.

To address these challenges, in this paper, we introduce the `Sica` (**S**emanti**c**ally Co**m**plementary Spectr**a**l Views). Instead of overlapping augmentations, `Sica` generates a pair of complementary spectral views of a graph: a *low-frequency view* that captures low-frequency signals related to community structure and homophily, and a *high-frequency view* that emphasizes high-frequency variations indicative of heterophily and local dissimilarities. Inspired by the Multi-View Information Bottleneck (MIB) principle [6, 19, 45], we design a special extractor with *score reconstruction attention (SRA)* to extract the most informative and compact substructure from each view, effectively filtering out noise. To further globally refine the spectral information from these distilled substructures, we introduce an *anchor-guided condensation mechanism*, condensing the global patterns of each subgraph into a single summary embedding. Finally, we perform hierarchical contrastive learning across four granularities: node, graph, core and anchor, to maximize mutual information. This progressive design not only strengthens alignment across all scales, from features to the global

essence of graph, but also amplifies subtle anomalous signals that traditional methods often overlook.

- ***New paradigm.*** We introduce a concept of semantically complementary spectral views for graph-level anomaly detection, capturing complementary *low-frequency and high-frequency information* to provide informative representations that are more conducive to anomaly detection.
- ***Novel Algorithm.*** We innovatively design a special extractor equipped with a score reconstruction attention and an anchor node-guided condensation mechanism. It can eliminate view-specific noise in complementary views by *local distillation and global condensation*, capture shared information, and extract compact spectral representations.
- ***Extensive experiments.*** We conduct comprehensive experiments on multiple benchmark datasets for *both graph-level anomaly detection and out-of-distribution detection tasks*, which demonstrate that `Sica` significantly outperforms state-of-the-art baselines in average performance.

## 2 Related Work

### 2.1 Graph-Level Anomaly Detection

The task of graph-level anomaly detection (GLAD) [35, 23, 42] aims to identify graphs that exhibit unusual patterns compared to a larger corpus. Early approaches often relied on graph kernels [34, 26] or handcrafted features to project graphs into a vector space, where classical anomaly detectors like One-Class SVM (OCSVM) [2] or Isolation Forest [15] could be applied. While effective to some extent, these methods are limited by their dependency on feature engineering and their inability to capture complex, non-linear structural patterns [14].

The advent of deep learning, particularly Graph Neural Networks (GNNs), has shifted the paradigm towards end-to-end representation learning for GLAD. Initial deep approaches often adapted autoencoder-based architectures, training a GNN to reconstruct normal graphs and using the reconstruction error as an anomaly score [21]. More recently, the field has been dominated by the principles of semi-supervised or self-supervised (e.g., Graph Contrastive Learning) [28, 19, 13, 20, 48, 11, 3], which have demonstrated superior performance by learning more discriminative representations without explicit reconstruction.

### 2.2 Contrastive Learning & Information Theory

Graph Contrastive Learning (GCL) [8, 37, 29] serves as the predominant framework for unsupervised GLAD. From an information-theoretic perspective, the goal of GCL is to learn a representation (or encoder) that maximizes the mutual information (MI) between representations of different augmented views of the same graph, a quantity for which the popular InfoNCE loss is a tractable lower bound [4]. The underlying assumption is that an ideal encoder should be invariant to nuisance variations introduced by augmentation, while retaining the core semantic information of the graph.

However, the efficacy of this approach is undermined by a fundamental flaw in existing view generation strategies. Prevailing methods, whether based on random perturbations [46, 33] or deterministic transformations [18, 9, 10, 19, 45], produce views that are informationally overlapping and semantically homogeneous. This high redundancy trivializes the contrastive objective, allowing the model to achieve low loss by learning superficial correlations—a form of representation collapse—rather than discovering the graph's core semantics. Crucially for GLAD, this flaw means anomalous patterns are often mirrored across views, resulting in a deceptively high consistency score that masks the abnormality. Our work contrasts semantically complementary spectral views to create a deliberately challenging task, compelling the model to discard superficial correlations and learn abstract representations fundamentally more sensitive to anomalies.

## 3 Methodology

### 3.1 Problem Definition

In this work, we focus on ***unsupervised*** graph-level anomaly detection. Given a dataset $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$, our learning goal is to train an anomaly detector $f(\cdot)$ on $\mathcal{D}_{train} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$ without any labels. During inference on a test set $\mathcal{D}_{test} = \{\mathcal{G}'_1, \mathcal{G}'_2, \ldots, \mathcal{G}'_m\}$ containing both normal

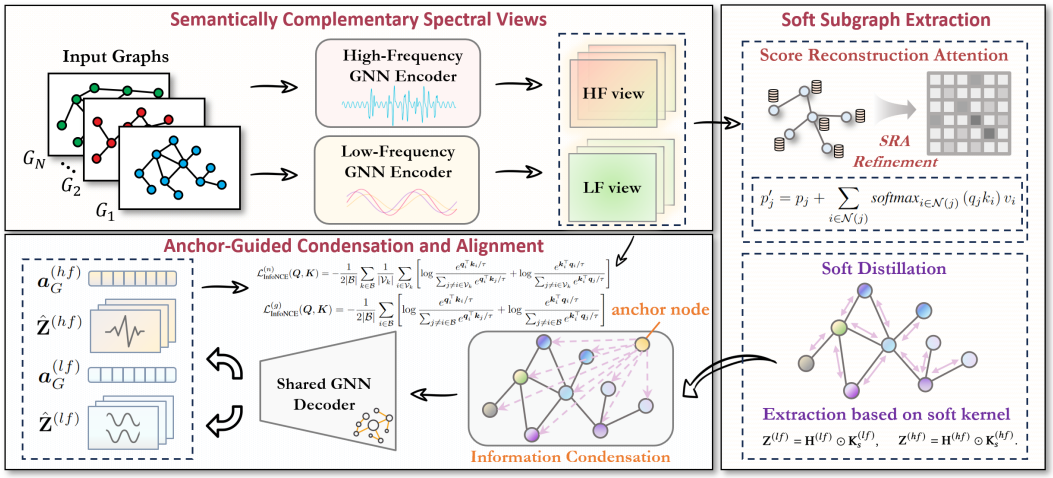



Figure 2: Architecture overview of the proposed **Sica**. Semantically complementary spectral views are generated from the input graph and distilled into informative subgraphs representations by SRA and anchor. A contrastive objective then aligns these representations, using the final consistency loss as the anomaly score for inference.

and anomalous graphs, a higher score $s_i = f(\mathcal{G}_i')$ indicates a higher probability that graph $\mathcal{G}_i'$ is anomalous. Formally, the anomaly detection task can be defined as:

$$\Psi\left(\mathcal{G}_i';\tau,f\right)=\begin{cases}1\,(\textit{abnormal}) & \text{if } f\left(\mathcal{G}_i'\right)>\tau\\ 0\,(\textit{normal}) & \text{otherwise}\end{cases} \tag{1}$$

where $\tau$ denotes the threshold and $\Psi\left(\cdot\right)$ is the decision function.

### 3.2 Overview

In this section, we detail the proposed **Sica** for GLAD. The core principle of **Sica** is to move beyond conventional homogeneous augmentations and instead learn from semantically complementary views derived from the spectral properties. As shown in Figure 2, the pipeline consists of four key stages. ① First, we generate two complementary views, capturing the low-frequency and high-frequency information of the graph, respectively (Sec 3.3). ② Second, inspired by the Multi-view Information Bottleneck principle, we design a soft subgraph extraction module with Score Reconstruction Attention to distill the most informative representation from each noisy view (Sec 3.4). ③ Third, we introduce an anchor-guided condensation mechanism, which further purifies the distilled subgraph and condenses it core semantics into a single anchor vector (Sec 3.5). ④ Finally, we employ contrastive objective to enforce cross-view alignment and optimize the information bottleneck objective, thereby amplifying subtle anomalous signals (Sec 3.6).

### 3.3 Semantically Complementary Spectral Views

A fundamental limitation of existing Graph Contrastive Learning frameworks is their reliance on generating semantically homogeneous views, typically through random perturbations, parallel transformations of features and structures, or uniform low-pass filtering. This paradigm creates homogeneous views that capture highly overlapping semantic information, trivializing the contrastive objective and effectively masking the subtle, high-frequency signals often indicative of anomalies. To overcome this, **Sica** abandons homogeneous augmentations in favor of a principled approach that constructs two ***semantically complementary*** views directly from spectral properties: ❖ a **Low-Frequency View** to capture stable, homophilous patterns and global community structure, and ❖ a **High-Frequency View** to highlight volatile, heterophilous signals and local node dissimilarities.

**Low-Frequency View Encoder ($\mathcal{F}_{lf}$) .** To capture low-frequency information, we employ a GNN that functions as a natural low-pass filter. Specifically, we select the *Graph Isomorphism Network (GIN)* [44], whose expressive power is equivalent to the Weisfeiler-Lehman (WL) test. Its message-passing mechanism excels at aggregating local neighborhood information, making it ideally suited for

learning representations of the community structures that define the low-frequency essence. Formally, given an input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$, the representation $\mathbf{h}_i^{(lf,\,l)}$ for node $v_i$ at layer $l$ of our low-pass encoder $\mathcal{F}_{lf}$ is computed as:

$$\mathbf{h}_i^{(lf,\,l)} = \text{MLP}_{lf}^{(l)}\big((1+\epsilon)\mathbf{h}_i^{(lf,\,l-1)} + \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j^{(lf,\,l-1)}\big), \tag{2}$$

where $\mathbf{h}_i^{(lf,\,0)} = \mathbf{x}_i$ are the initial node features, $\mathcal{N}(i)$ is the neighborhood of node $v_i$, and $\epsilon$ is a learnable parameter. The final node representations for this view are obtained by concatenating the outputs from all $L$ layers:

$$\mathbf{h}_i^{(lf)} = \Big[\mathbf{h}_i^{(lf,\,1)} \,\|\, \mathbf{h}_i^{(lf,\,2)} \,\|\, \cdots \,\|\, \mathbf{h}_i^{(lf,\,L)}\Big]. \tag{3}$$

**High-Frequency View Encoder ($\mathcal{F}_{hf}$) .** While the low-frequency view captures the graph's stable backbone, it is inherently insensitive to the high-frequency aberrations that often signify anomalies. Therefore, we introduce a complementary high-frequency view, generated using a high-pass spectral filter. We leverage the *Beta Wavelet Graph Neural Network (BWGNN)* [30], whose spectrally localized wavelet kernels can be configured to respond specifically to high-frequency components in the graph signal. This enables `Sica` to explicitly model sharp variations between nodes and long-range, heterophilous connections, which low-pass GNNs would smooth over and ignore. BWGNN builds upon Beta distribution-based wavelet kernels $\mathcal{W}_{p,q}$. The propagation rule is given by:

$$\mathcal{W}_{p,q} = \mathbf{U}\beta_{p,q}^{*}(\mathbf{\Lambda})\mathbf{U}^T = \frac{\left(\frac{L}{2}\right)^p \left(I - \frac{L}{2}\right)^q}{2B(p+1, q+1)}, \tag{4}$$

where $C = p + q$ is a constant, and $B\,(p+1, q+1) = \frac{p!\,q!}{(p+q+1)!}$. Then we can get a set of different Beta wavelets kernels with the same order $\{\mathcal{W}_{0,\,C}, \mathcal{W}_{1,\,C-1}, \cdots, \mathcal{W}_{C,\,0}\}$. The propagation rule of BWGNN can be expressed as:

$$\mathbf{h}_i^{(hf,k)} = \mathcal{W}_{k,\,C-k}\left(\text{MLP}\left(\mathbf{x}_i\right)\right), \tag{5}$$

where $\mathcal{W}_{k,\,C-k}$ is the $k$-th wavelet kernel designed to capture a specific frequency band. By selecting kernels corresponding to higher frequencies, $\mathcal{F}_{hf}$ effectively extracts the desired signals. The final node representations for the high-frequency view are an aggregation of outputs from different kernels:

$$\mathbf{h}_i^{(hf)} = \Big[\mathbf{h}_i^{(hf,\,1)} \,\|\, \mathbf{h}_i^{(hf,\,2)} \,\|\, \cdots \,\|\, \mathbf{h}_i^{(hf,\,L)}\Big]. \tag{6}$$

**Graph-Level Readout.** Given the obtained node-level representations $\mathbf{H}^{(lf)}$ and $\mathbf{H}^{(hf)}$ from both views, we apply mean pooling to obtain graph-level representations for contrastive learning:

$$\mathbf{h}_G^{(lf)} = \frac{1}{|\mathcal{V}|}\sum_{i\,\in\,\mathcal{V}} \mathbf{h}_i^{(lf)}, \quad \mathbf{h}_G^{(hf)} = \frac{1}{|\mathcal{V}|}\sum_{i\,\in\,\mathcal{V}} \mathbf{h}_i^{(hf)}, \tag{7}$$

where $\mathbf{h}_G^{(lf)}$ and $\mathbf{h}_G^{(hf)}$ denote the graph-level features of two views.

### 3.4 Soft Subgraph Extraction for Local Information Distillation

Complementary spectral views capture distinct anomalous signals, but the low-frequency view contains redundant global background and the high-frequency view suffers from local noise. Direct contrastive learning on these views would be dominated by noise. Inspired by MIB [6], we use local distillation to extract informative subgraphs, learning subgraph representations $\mathbf{Z}$ that maximize cross-view shared information while minimizing view-specific redundancy. Formally, we define a symmetric optimization objective for views $\mathbf{H}^{(lf)}$ and $\mathbf{H}^{(hf)}$. For $\mathbf{H}^{(lf)}$, we have

$$\min_{\mathbf{Z}^{(lf)}} \quad \underbrace{\mathcal{I}(\mathbf{H}^{(lf)}; \mathbf{Z}^{(lf)} | \mathbf{H}^{(hf)})}_{\text{view-specific redundancy}} - \lambda \underbrace{\mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{H}^{(hf)})}_{\text{shared information}}, \tag{8}$$

where $\mathcal{I}(A;B|C)$ is conditional mutual information and $\lambda$ is a trade-off hyperparameter. This symmetric objective can be simplified to a unified information bottleneck objective:

$$\mathcal{L}_{\text{distill}} = \underbrace{\mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)})}_{\text{Compression}} - \gamma \underbrace{\mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)})}_{\text{Informativeness}}, \tag{9}$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy and $\mathcal{H}(\mathcal{G}_s)$ measuring the structural randomness of the subgraph $\mathcal{G}_s$, and $\gamma$ is a reparameterized balancing hyperparameter. We defer the detailed derivation of Eq. (8) and Eq. (9) to Appendix B.

This objective reveals a dual requirement for effective subgraph extraction: ***maximize cross-view shared information*** and ***minimize subgraph structural randomness***. Existing subgraph extraction methods usually only optimize the mutual information term, ignoring the entropy term, which leads to subgraphs with lost structural information and semantic defects. To address this, we propose a extraction pipeline:

**Raw Score Generator.** We first adopt two independent subgraph extractors $e_{lf}(\cdot \mid \mathbf{H}^{(lf)})$ and $e_{hf}(\cdot \mid \mathbf{H}^{(hf)})$ to obtain compact representations for views. Each extractor outputs a node importance vector $\boldsymbol{p} \in \mathbb{R}^{|\mathcal{V}|\times 1}$, where $p_i$ is the probability of node $v_i$ being selected into the core subgraph.

**Score Reconstruction Attention.** To encourage the extraction of cohesive substructures, we adopt a *Score Reconstruction Attention (SRA)* mechanism to reconstruct the raw node importance scores. Formally, we define the query, key, and value uniformly: $q_j = p_j, k_i = p_i, v_i = p_i$. The reconstructed importance score $p'_j$ is then obtained by a residual connection to preserve the original score:

$$p'_j = p_j + \sum_{i \in \mathcal{N}(j)} \textit{softmax}_{i\in\mathcal{N}(j)}\left(q_j k_i\right) v_i. \tag{10}$$

This propagates importance signals across edges, naturally biasing substructures selection towards connected components. This implicitly satisfies the entropy minimization requirement of Eq. (9).

**Soft Distillation.** The reconstructed importance scores vector $\boldsymbol{p}' \in \mathbb{R}^{|\mathcal{V}|\times 1}$ are broadcasted to construct the soft kernel $\mathbf{K}_s \in \mathbb{R}^{|\mathcal{V}|\times d}$. We then perform feature soft distillation via Hadamard product:

$$\mathbf{Z}^{(lf)} = \mathbf{H}^{(lf)} \odot \mathbf{K}_s^{(lf)}, \quad \mathbf{Z}^{(hf)} = \mathbf{H}^{(hf)} \odot \mathbf{K}_s^{(hf)}. \tag{11}$$

Therefore, $\mathbf{Z}^{(lf)}$ and $\mathbf{Z}^{(hf)}$ adaptively attenuate irrelevant node features and softly distill the core subgraph semantics, avoiding rigid node dropping.

### 3.5 Anchor-Guided Condensation for Global Information Condensation

Similar to the core objective of local distillation. In this section, we extend Eq.(9) to the global representation space and condense the scattered core subgraph information into a unified global representation $\mathbf{a}_G$. Therefore, the global objective is defined as:

$$\mathcal{L}_{\text{global}} = \underbrace{\mathcal{H}(\boldsymbol{a}_G^{(lf)}) + \mathcal{H}(\boldsymbol{a}_G^{(hf)})}_{\text{Global Compression}} - \gamma \underbrace{\mathcal{I}(\boldsymbol{a}_G^{(lf)}; \boldsymbol{a}_G^{(hf)})}_{\text{Global Informativeness}} \tag{12}$$

where $\gamma$ is a hyperparameter balancing compression and informativeness at the global stage.

**Anchor-Node as Information Sink.** To distill a compact and decisive representation of the graph's absolute core, we propose an *anchor-guided condensation mechanism*, which aggregates soft-distilled node representations into a single robust ***anchor node***. Formally, We introduce a virtual node $v_a$, initialized with a zero feature vector and connected to all nodes in $\mathcal{V}$, to form an anchor-guided core graph $\tilde{\mathcal{G}} = (\mathcal{V} \cup \{v_a\}, \mathcal{E} \cup \{(v_a, v) \mid v \in \mathcal{V}\})$, with node representations $\tilde{\mathbf{Z}} = [\mathbf{Z}; \mathbf{x}_a]$, where $\mathbf{Z} \in \{\mathbf{Z}^{(lf)}, \mathbf{Z}^{(hf)}\}$. This anchor-guided graph $\tilde{\mathcal{G}}$ is processed by a band-pass decoder $\mathcal{F}_\theta$. In our practice, $\mathcal{F}_\theta$ is instantiated as a BWGNN, whose parameters are shared across views:

$$\hat{\mathbf{Z}}^{(lf)} = \mathcal{F}_\theta\left(\tilde{\mathbf{Z}}^{(lf)}, \tilde{\mathbf{A}}\right), \quad \hat{\mathbf{Z}}^{(hf)} = \mathcal{F}_\theta\left(\tilde{\mathbf{Z}}^{(hf)}, \tilde{\mathbf{A}}\right). \tag{13}$$

Within this architecture, the anchor node serves as an ***information sink***, aggregating messages from all core nodes via message passing. The resulting anchor representations $\boldsymbol{a}_G^{(lf)}$ and $\boldsymbol{a}_G^{(hf)}$ (i.e., the

last rows of $\hat{\mathbf{Z}}^{(lf)}$ and $\hat{\mathbf{Z}}^{(hf)}$) provide compact summaries of the distilled semantics for each view, while the remaining rows correspond to the updated core node representations (i.e., $\hat{\mathbf{z}}_i^{(lf)}$ and $\hat{\mathbf{z}}_i^{(hf)}$).

For the global compression constraint of Eq. (12), this process compresses scattered node-level representations into a single vector, inherently reducing the entropy of global representations.

### 3.6 Alignment for Mutual Information Maximization

In practice, directly optimizing the mutual information term in the information bottleneck objective is intractable for high-dimensional graph representations. We therefore adopt the InfoNCE loss [27] as a tractable lower bound approximation. See Appendix B for more details.

Single-level contrastive learning fails to capture anomalous patterns emerging across diverse structural granularities. Therefore, we design a hierarchical contrastive framework that enforces cross-view alignment at four levels: node, graph, core, and anchor. By modeling multi-scale consistency of normal graphs, `Sica` amplifies structural deviations for accurate anomaly detection. We define two generic InfoNCE losses (node-level and graph-level) for representation pairs $(\boldsymbol{Q}, \boldsymbol{K})$:

$$\mathcal{L}_{\text{InfoNCE}}^{(n)}(\boldsymbol{Q}, \boldsymbol{K}) = -\frac{1}{2|\mathcal{B}|} \sum_{k \in \mathcal{B}} \frac{1}{|\mathcal{V}_k|} \sum_{i \in \mathcal{V}_k} \left[ \log \frac{e^{\boldsymbol{q}_i^\top \boldsymbol{k}_i / \tau}}{\sum_{j \neq i \in \mathcal{V}_k} e^{\boldsymbol{q}_i^\top \boldsymbol{k}_j / \tau}} + \log \frac{e^{\boldsymbol{k}_i^\top \boldsymbol{q}_i / \tau}}{\sum_{j \neq i \in \mathcal{V}_k} e^{\boldsymbol{k}_i^\top \boldsymbol{q}_j / \tau}} \right] \tag{14}$$

$$\mathcal{L}_{\text{InfoNCE}}^{(g)}(\boldsymbol{Q}, \boldsymbol{K}) = -\frac{1}{2|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[ \log \frac{e^{\boldsymbol{q}_i^\top \boldsymbol{k}_i / \tau}}{\sum_{j \neq i \in \mathcal{B}} e^{\boldsymbol{q}_i^\top \boldsymbol{k}_j / \tau}} + \log \frac{e^{\boldsymbol{k}_i^\top \boldsymbol{q}_i / \tau}}{\sum_{j \neq i \in \mathcal{B}} e^{\boldsymbol{k}_i^\top \boldsymbol{q}_j / \tau}} \right] \tag{15}$$

All losses are instantiated from the above forms. Specifically, we define $\mathcal{L}_s = \mathcal{L}_{\text{InfoNCE}}^{(f)}\big(\mathbf{c}^{(lf)}, \mathbf{c}^{(hf)}\big)$, where $s \in \{\text{node}, \text{graph}, \text{core}, \text{anchor}\}$ denotes the optimization level, $f \in \{n, g\}$ selects the InfoNCE variant. The representation $\mathbf{c}$ takes the form of $\mathbf{H}$, $\mathbf{h}_G$, $\hat{\mathbf{Z}}$, and $\mathbf{a}_G$ corresponding to the node, graph, core, and anchor levels, respectively. Notably, $\mathcal{L}_{\text{core}}$ and $\mathcal{L}_{\text{anchor}}$ serve as the primary terms to explicitly optimize the *mutual information terms* in Eq. (9) and Eq. (12). $\mathcal{L}_{\text{node}}$ and $\mathcal{L}_{\text{graph}}$ impose feature-level regularization to preserve the spectral properties of original views, further facilitating mutual information maximization. The *compression terms* in Eq. (9) and Eq. (12) are implicitly optimized via the SRA and anchor mechanism. For more detailed discussions, refer to Appendix B. Thus, we form the final training loss:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{node}} + \mathcal{L}_{\text{graph}}}_{\text{regularization}} + \underbrace{\alpha(\mathcal{L}_{\text{core}} + \mathcal{L}_{\text{anchor}})}_{\text{mutual information}}, \tag{16}$$

where $\alpha$ denotes a hyperparameter balancing the mutual information and regularization terms.

During inference, we use the contrastive objective as the anomaly score. Normal graphs exhibit high cross-view consistency and thus yield low loss values, whereas anomalous graphs induce semantic misalignment across views, resulting in higher losses. The anomaly score for a graph $\mathcal{G}$ is defined as:

$$s(\mathcal{G}) = s_{\text{node}} + s_{\text{graph}} + \beta\big(s_{\text{core}} + s_{\text{anchor}}\big), \tag{17}$$

where each $s_{(\cdot)}$ denotes the corresponding contrastive loss and $\beta$ denotes a balancing hyperparameter.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.**We evaluate `Sica` on both graph-level anomaly detection (GLAD) and graph-level out-of-distribution (GLOD) detection benchmarks provided by UB-GOLD [38], covering diverse domains such as *social networks, protein interaction networks, and molecular graphs*. For GLAD, we adopt 10 widely used benchmark datasets (e.g., ENZYMES, PROTEINS) from the TU benchmark [25] and Tox21 challenge [1]. Following the standard UB-GOLD protocol, anomalies are either inherently present (Tox21) or defined as samples from minority or distinct classes (TU datasets). For GLOD, we evaluate on 6 benchmark dataset pairs with inter-dataset shifts (e.g., AIDS→DHFR, BZR→COX2) from UB-GOLD [18], where one dataset serves as in-distribution (ID) and the other as out-of-distribution (OOD). Detailed statistics are provided in Appendix D.

Table 1: Performance comparison (AUROC %, mean±std) of different methods on GLAD benchmark datasets. The best results are highlighted in **bold**, and the second best are underlined.

| Dataset | Sica | OCGIN | OCGTL | SIGNET | CVTGAD | MUSE | GLADPro | GOOD-D | SEGO | GOODAT |
|---|---|---|---|---|---|---|---|---|---|---|
| MMP | **71.41±1.00** | 69.37±3.46 | 67.51±5.20 | 71.23±4.24 | 70.58±4.69 | 63.48±10.04 | 53.01±9.39 | 71.41±7.14 | - | 67.49±4.24 |
| IMDB-B | 65.73±3.87 | 57.90±4.69 | 69.10±1.73 | 69.07±3.61 | 65.36±3.46 | **73.54±2.84** | 62.36±14.02 | 66.29±3.46 | 64.13±3.09 | 65.46±9.17 |
| REDDIT-B | 85.77±1.73 | 82.10±6.08 | **89.92±3.46** | 85.24±6.71 | 87.43±7.75 | 83.62±4.51 | 66.13±22.12 | 89.43±4.69 | - | 80.31±0.77 |
| ENZYMES | **66.44±5.43** | 62.44±6.16 | 63.59±3.32 | 63.12±7.21 | 65.12±7.21 | 50.80±16.33 | 52.08±4.13 | 65.54±4.47 | 65.37±7.87 | 52.33±7.68 |
| PROTEINS | **78.26±2.45** | 76.46±3.61 | 72.89±7.55 | 75.86±5.48 | 76.49±5.39 | 69.41±3.19 | 75.62±2.46 | 76.02±4.12 | 57.90±4.32 | 77.92±8.66 |
| DD | **79.12±3.16** | 79.08±4.36 | 77.76±6.93 | 72.30±3.87 | 77.36±6.32 | 78.62±4.14 | 70.46±4.20 | 76.62±3.61 | 56.15±5.35 | 77.62±5.10 |
| BZR | **84.27±6.37** | 69.13±3.61 | 51.89±6.78 | 80.79±6.16 | 77.69±5.29 | 68.25±9.67 | 63.61±10.83 | 73.28±3.16 | 76.03±10.30 | 64.77±3.46 |
| AIDS | **99.54±0.23** | 96.89±2.47 | 99.36±0.19 | 97.60±1.29 | 99.21±0.56 | 99.17±0.42 | 56.36±17.95 | 97.10±1.69 | 89.40±2.65 | 98.82±0.78 |
| COX2 | **74.83±2.20** | 57.81±7.07 | 59.81±5.48 | 72.35±3.62 | 64.36±3.12 | 53.33±1.34 | 57.24±7.14 | 63.19±4.57 | 63.20±6.32 | 59.99±4.47 |
| DHFR | **71.34±2.83** | 59.95±3.00 | 59.62±3.32 | 70.45±3.60 | 69.31±3.32 | 65.36±10.45 | 55.72±5.92 | 68.94±1.73 | 65.68±3.36 | 61.52±5.32 |
| Avg. AUROC | **77.67** | 71.11 | 71.14 | 75.8 | 75.29 | 70.56 | 61.26 | 74.78 | - | 70.62 |
| Avg. Rank | **1.7** | 6.0 | 5.0 | 4.2 | 3.8 | 6.0 | 8.4 | 4.0 | - | 5.8 |

Table 2: Performance comparison (AUROC %, mean±std) of different methods on GLOD benchmark datasets. The best results are highlighted in **bold**, and the second best are underlined.

| Dataset | Sica | OCGIN | OCGTL | SIGNET | CVTGAD | MUSE | GLADPro | GOOD-D | SEGO | GOODAT |
|---|---|---|---|---|---|---|---|---|---|---|
| IM→IB | **84.27±2.65** | 80.98±5.57 | 66.73±7.21 | 71.10±7.94 | 80.23±5.83 | 79.50±3.66 | 72.06±4.02 | 80.94±4.12 | 78.17±1.73 | 77.66±7.14 |
| EN→PR | **68.24±3.00** | 61.77±5.92 | 67.18±7.81 | 62.42±5.48 | 64.31±7.48 | 64.12±4.27 | 62.17±4.52 | 63.84±5.29 | 65.08±4.46 | 64.61±6.78 |
| AI→DH | **99.91±0.06** | 95.68±1.98 | 98.95±0.81 | 96.82±2.08 | 99.10±0.46 | 99.13±0.36 | 79.64±16.94 | 99.27±0.24 | 98.68±0.59 | 93.95±4.66 |
| BZ→CO | **99.31±0.89** | 87.27±5.66 | 81.86±5.39 | 89.11±6.08 | 96.32±2.12 | 74.87±6.71 | 69.79±7.41 | 95.16±3.13 | 92.29±1.33 | 80.97±5.10 |
| BB→BA | **85.53±2.00** | 78.29±2.00 | 78.89±2.24 | 85.45±1.98 | 84.51±2.00 | 80.66±2.19 | 55.44±5.56 | 81.44±3.32 | 84.80±1.94 | 75.94±3.61 |
| PT→MU | **92.11±1.93** | 74.07±7.42 | 67.51±10.20 | 85.22±2.35 | 79.78±4.24 | 85.09±3.47 | 69.53±18.06 | 85.06±2.83 | 86.01±2.10 | 80.42±4.90 |
| Avg. AUROC | **88.23** | 79.68 | 76.85 | 81.69 | 84.04 | 80.56 | 68.11 | 84.29 | 84.17 | 78.92 |
| Avg. Rank | **1.0** | 7.0 | 6.83 | 5.67 | 4.33 | 5.5 | 9.33 | 4.17 | 4.0 | 7.17 |

**Baseline Methods.** We compare `Sica` with baseline methods from two categories: ✤ *end-to-end GLAD methods*, including OCGIN [48], OCGTL [28], CVTGAD [13], SIGNET [19], MUSE [11] and GLADPro [45]; ✤ *end-to-end GLOD methods*, including GOOD-D [18], SEGO [9] and GOODAT [36]. We provide more introduction to baselines in Appendix E.

**Evaluation Metrics and Implementation.** We use the area under the ROC curve (**AUROC**) as the primary evaluation metric for both GLAD and GLOD tasks, which provides a threshold-independent measure of detection performance with higher values indicating better performance. We repeat all experiments 5 runs, reporting the mean AUROC with standard deviation, and summarize overall performance by average rank across datasets.

### 4.2 Anomaly Detection Results

We evaluate `Sica` on benchmark GLAD datasets, the results are reported in Table 1. On 8 out of the 10 datasets, `Sica` consistently achieves the state-of-the-art performance and attains the highest average performance among all competing methods. This indicates strong and stable anomaly detection capability. In particular, `Sica` exhibits strong performance on *proteins and molecules* datasets. For examples, on BZR and COX2, `Sica` improves upon the strongest baseline by 4% and 2% AUROC, respectively. Anomalies in these domains often correspond to subtle yet functionally critical deviations in global topology or key substructures rather than isolated node-level abnormalities. By aligning complementary spectral views, `Sica` is well suited to capture such global semantic inconsistencies. On *social network* datasets such as IMDB-B and REDDIT-B, despite the fact that the graphs are typically larger and noisier, and anomalies are not clearly defined by structural semantics, `Sica` still achieves competitive performance and consistently ranks highly. These results demonstrate that `Sica` can effectively balance sensitivity to complex structural anomalies with generalization ability to heterogeneous graph domains, making it a reliable GLAD framework for practical applications.

### 4.3 Out-of-distribution Detection Results

We further evaluate `Sica` on 6 GLOD benchmark datasets to validate whether the representations learned by `Sica` generalize well to GLOD tasks. The results are presented in Table 2. Notably, `Sica` exhibits clear superiority on all OOD dataset pairs, including *social network* graph scenarios. For instance, on challenging OOD pairs IM→IB, BZ→CO and PT→MU, `Sica` surpasses the strongest baseline by 2%, 3% and 7% in AUROC, respectively, which demonstrates its powerful generalization capability. This indicates that although `Sica` excels at capturing fine-grained anomalies in structured domains such as *molecular graphs*, its spectral representations are particularly suitable for identifying distributional boundaries, enabling strong performance in GLOD across diverse domains. We have discussed more details about the advantages of `Sica` in the Appendix G.

### 4.4 Ablation Study

We conduct ablation experiments from two different perspectives: ***Model Components*** and ***Loss Functions***, to verify the contributions of paradigms and modules in `Sica`. In particular, the ablation on loss functions also provides indirect evidence for the necessity of our proposed components.

**Studies on Model Components.** Table 3 reports the results of ablating key components in `Sica`. Specifically, discarding the complementary views and relying solely on the uniform low-pass filtering leads to a substantial performance drop, as the resulting information exhibit highly overlapping views collapse and limited discriminability. Removing SRA causes a moderate drop, this indicates that SRA can effectively encourage the extraction of cohesive substructures and handle the structures randomness. Among all variants, removing the anchor-guided condensation mechanism results in a severe performance degradation. Especially on challenging OOD pairs such as PT→MU, the AUROC drops below 80%. Interestingly, replacing the band-pass GNN decoder with a low-pass GNN decoder further degrades performance, verifying that mid- or high-frequency graph signals are essential for anchor aggregation. This confirms that each component plays a central role in `Sica`. More comprehensive ablation results on model components across additional datasets are provided in the Appendix F.

Table 3: Ablation study of `Sica` on benchmarks (AUROC %, mean). Each component is removed to verify its contribution.

| Method | BZR | COX2 | PT→MU | EN→PR | Avg. |
|---|---|---|---|---|---|
| `Sica` (w/o **spectral views**) | 79.12 | 72.21 | 87.28 | 64.03 | 75.66 |
| `Sica` (w/o **band-pass decoder**) | 73.25 | 67.58 | 78.82 | 64.24 | 70.97 |
| `Sica` (w/o **SRA**) | 80.44 | 72.67 | 91.57 | 64.07 | 77.19 |
| `Sica` (w/o **anchor node**) | 70.24 | 65.45 | 77.90 | 57.98 | 67.89 |
| `Sica` (full) | **84.27** | **74.83** | **92.11** | **68.24** | **79.86** |

**Studies on Loss Functions.** Table 4 reports the effect of loss terms. Experimental results show that $\mathcal{L}_{anchor}$ is the most important term: removing it leads to a sharp AUROC drop, which verifies that global cross-view alignment is crucial. Eliminating $\mathcal{L}_{graph}$ also causes a noticeable AUROC drop, demonstrating that global feature regularization is essential. In contrast, removing $\mathcal{L}_{node}$ or $\mathcal{L}_{core}$ only leads to a moderate performance degradation, yet the corresponding optimization is also indispensable. More importantly, we observe a clear *coupling effect* between $\mathcal{L}_{core}$ and $\mathcal{L}_{anchor}$. The supervision of $\mathcal{L}_{core}$ brings limited performance gains without $\mathcal{L}_{anchor}$ support, while it can further boost detection performance when $\mathcal{L}_{anchor}$ is activated. This validates that global alignment is enhances the utility of SRA supervision, and the full combination of all loss terms jointly achieves optimal overall performance.

Table 4: Ablation study of `Sica` on benchmarks (AUROC %, mean). Each loss is removed to verify its contribution.

| $\mathcal{L}_{node}$ | $\mathcal{L}_{graph}$ | $\mathcal{L}_{core}$ | $\mathcal{L}_{anchor}$ | BZR | COX2 | PT→MU | EN→PR | Avg. |
|---|---|---|---|---|---|---|---|---|
| ✗ | ✓ | ✓ | ✓ | 82.69 | 72.20 | 90.92 | 65.97 | 77.95 |
| ✓ | ✗ | ✓ | ✓ | 79.34 | 70.28 | 86.88 | 63.94 | 75.11 |
| ✓ | ✓ | ✗ | ✓ | 80.65 | 70.15 | **93.34** | 63.56 | 76.93 |
| ✓ | ✓ | ✓ | ✗ | 71.72 | 65.37 | 74.78 | 62.78 | 68.66 |
| ✓ | ✓ | ✗ | ✗ | 72.97 | 65.89 | 76.84 | 63.74 | 69.86 |
| ✓ | ✓ | ✓ | ✓ | **84.27** | **74.83** | 92.11 | **68.24** | **79.86** |

### 4.5 Sensitivity Analysis

We analyze the sensitivity of `Sica` with respect to two key hyperparameters: the loss-balancing coefficient $\alpha$ and the scoring coefficient $\beta$. These two hyperparameters control the contribution of loss terms, and thus directly affect the final detection performance. Figure 3(a) reports the impact of $\alpha$ during training. `Sica` maintains stable performance over a wide range (e.g., $\alpha \in [0.4, 1.8]$), showing the training objective requires no delicate parameter tuning. Figure 3(b) reports the impact of $\beta$ during inference. This show a smooth trend, confirming that the anomaly score remains robust to moderate changes of $\beta$. Notably, performance drops significantly when $\alpha = 0.0$ or $\beta = 0.0$, as the corresponding loss term is disabled. This indirectly validates the importance of our introduced loss terms.

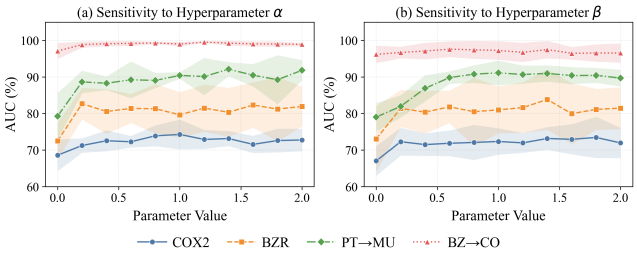



Figure 3: Sensitivity analysis of key balancing hyperparameters $\alpha$ and $\beta$ on benchmarks.

### 4.6 Robustness Analysis

**Noise Contamination** To evaluate whether `Sica` is robust to different levels of noise contamination, we inject *Gaussian perturbations* into node features during training, thereby simulating realistic scenarios in which node attributes may be noisy or partially corrupted. The results are reported in Figure 4. In setting (a), we gradually increase the *contamination intensity* $\sigma$ for all nodes. As $\sigma$ increases, the AUROC decreases only slightly within a reasonable range, indicating that `Sica` is resilient to feature noise. In setting (b), we fix $\sigma = 0.1$ and vary the *proportion* of perturbed nodes from 1% to 32%. `Sica` maintains stable performance across different ratios, indicating that it is tolerant to partial contamination. These results highlight the robustness of `Sica` under feature levels.

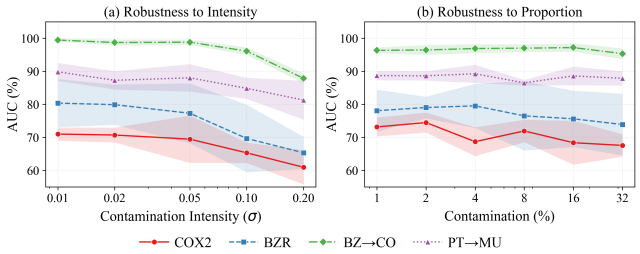



Figure 4: Robustness analysis of `Sica` w.r.t. noise contamination on benchmarks.

**Graph Contamination** To study robustness under contaminated training data, we inject varying proportions of anomalous graphs into the training set while keeping the model unaware and still assuming all training graphs are normal. As shown in Table 5, `Sica` remains stable on most datasets and even improves in some cases. Notably, REDDIT-B, despite its modest clean performance, achieves higher AUROC after contamination. These results indicate that `Sica` has good tolerance to a certain level of training graph contamination.

Table 5: AUROC (%) of `Sica` under different proportions of anomalous graphs injected into the training set.

| Dataset | 0% (clean) | 5% | 10% | 15% |
|---|---|---|---|---|
| MMP | **71.41** | 69.51 | 69.35 | 68.29 |
| AIDS | **99.54** | 97.90 | 98.03 | 95.00 |
| BZR | **84.27** | 79.15 | 78.15 | 78.55 |
| COX2 | **74.83** | 71.83 | 70.30 | 72.58 |
| DD | **79.12** | 77.31 | 77.09 | 76.92 |
| DHFR | **71.34** | 67.02 | 67.01 | 65.56 |
| ENZYMES | **66.44** | 61.42 | 64.70 | 63.98 |
| PROTEINS | 78.26 | 77.32 | 78.89 | **78.92** |
| IMDB-B | **65.73** | 61.09 | 50.64 | 50.68 |
| REDDIT-B | 85.77 | **87.49** | 87.25 | 87.23 |

## 5 Conclusion

In this paper, we propose `Sica` for graph-level anomaly detection. `Sica` replaces the semantically homogeneous views in existing methods with semantically complementary spectral views, which are more capable of capturing anomalous signals and help prevent shortcuts in contrastive learning. To further eliminate view-specific noise in these complementary views, we design a subgraph extractor with score reconstruction attention and an anchor-guided condensation mechanism to distill the essential shared information. Finally, we employ contrastive learning on multi-granularity representations to learn discriminative graph representations. Extensive experiments verify the effectiveness of `Sica` in graph-level anomaly detection and graph-level out-of-distribution detection.

## Acknowledgments and Disclosure of Funding

This work was supported in part by the Beijing Municipal Natural Science Foundation–Fengtai Innovation Joint Fund (Grant No. L2601032), the National Natural Science Foundation of China (Grant Nos. 62202042, 6212780016, and 624B2027), the State Key Laboratory of Internet Architecture, Tsinghua University (Grant No. HLW2025MS03), the Fundamental Research Funds for the Central Universities (Grant No. 2024JBMC031), the Aeronautical Science Foundation of China (Grant No. ASFC-2024Z0710M5002), the Open Fund of the Key Laboratory of Advanced Cryptography and System Security of Sichuan Province (Grant No. SKLACSS-202312), the Guangdong Science and Technology Programme (Grant No. 2024B0101030002), the Ministry of Industry and Information Technology of China, the National Key Research and Development Program of China (Grant No. 2023YFB3307500), the Science and Technology Innovation Project of Hunan Province (Grant No. 2023RC4014), the Open Project of the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences) (Grant No. 2024PY029), and the Fundamental Research Funds for the Central Universities, Jilin University.

## References


[1] Ahmed Abdelaziz, Hilde Spahn-Langguth, Karl-Werner Schramm, and Igor V Tetko. Consensus modeling for hts assays using in silico descriptors calculates the best balanced accuracy in tox21 challenge. *Frontiers in Environmental Science*, 4:2, 2016.

[2] Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 8–15, 2013.

[3] Jinyu Cai, Yunhe Zhang, Pengyang Wang, and See-Kiong Ng. Moegad: A mixture-of-experts framework with pseudo-anomaly generation for graph-level anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[5] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[6] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

[7] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[8] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, pages 4116–4126. PMLR, 2020.

[9] Yue Hou, He Zhu, Ruomei Liu, Yingke Su, Jinxiang Xia, Junran Wu, and Ke Xu. Structural entropy guided unsupervised graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17258–17266, 2025.

[10] Wangyu Jin, Huifang Ma, Yingyue Zhang, Zhixin Li, and Liang Chang. Dual-view graph-of-graph representation learning with graph transformer for graph-level anomaly detection. *Neural Networks*, 187: 107291, 2025.

[11] Sunwoo Kim, Soo Yong Lee, Fanchen Bu, Shinhwan Kang, Kyungho Kim, Jaemin Yoo, and Kijung Shin. Rethinking reconstruction-based graph-level anomaly detection: limitations and a simple remedy. *Advances in Neural Information Processing Systems*, 37:95931–95962, 2024.

[12] TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] Jindong Li, Qianli Xing, Qi Wang, and Yi Chang. Cvtgad: Simplified transformer with cross-view attention for unsupervised graph-level anomaly detection. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 185–200. Springer, 2023.

[14] Fanzhen Liu, Xiaoxiao Ma, Jia Wu, Jian Yang, Shan Xue, Amin Beheshti, Chuan Zhou, Hao Peng, Quan Z Sheng, and Charu C Aggarwal. Dagad: Data augmentation for graph anomaly detection. In *2022 IEEE international conference on data mining (ICDM)*, pages 259–268. IEEE, 2022.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[16] Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. Towards trustworthy rumor detection with interpretable graph structural learning. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4089–4093, 2023.

[17] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[18] Yixin Liu, Kaize Ding, Huan Liu, and Shirui Pan. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pages 339–347, 2023.

[19] Yixin Liu, Kaize Ding, Qinghua Lu, Fuyi Li, Leo Yu Zhang, and Shirui Pan. Towards self-interpretable graph-level anomaly detection. *Advances in Neural Information Processing Systems*, 36:8975–8987, 2023.

[20] Xuexiong Luo, Jia Wu, Jian Yang, Shan Xue, Hao Peng, Chuan Zhou, Hongyang Chen, Zhao Li, and Quan Z Sheng. Deep graph level anomaly detection with contrastive learning. *Scientific Reports*, 12(1): 19867, 2022.

[21] Rongrong Ma, Guansong Pang, Ling Chen, and Anton Van Den Hengel. Deep graph-level anomaly detection by glocal knowledge distillation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 704–714, 2022.

[22] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE transactions on knowledge and data engineering*, 35(12):12012–12038, 2021.

[23] Xiaoxiao Ma, Jia Wu, Jian Yang, and Quan Z Sheng. Towards graph-level anomaly detection via deep evolutionary mapping. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1631–1642, 2023.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[25] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

[26] Marion Neumann, Roman Garnett, Christian Bauckhage, and Kristian Kersting. Propagation kernels: efficient graph kernels from propagated information. *Machine learning*, 102(2):209–245, 2016.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[28] Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. Raising the bar in graph-level anomaly detection. *arXiv preprint arXiv:2205.13845*, 2022.

[29] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1150–1160, 2020.

[30] Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pages 21076–21089. PMLR, 2022.

[31] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

[32] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[33] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

[34] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

[35] Huawei Wang, Huamin Wang, Yijing Guo, and Zhong Li. A novel graph-level anomaly detection model. In *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*, pages 451–455, 2023.

[36] Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. Goodat: Towards test-time graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15537–15545, 2024.

[37] Yili Wang, Kaixiong Zhou, Rui Miao, Ninghao Liu, and Xin Wang. Adagcl: Adaptive subgraph contrastive learning to generalize large-scale graph training. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 2046–2055, 2022.

[38] Yili Wang, Yixin Liu, Xu Shen, Chenyu Li, Kaize Ding, Rui Miao, Ying Wang, Shirui Pan, and Xin Wang. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark. *arXiv preprint arXiv:2406.15523*, 2024.

[39] Yingheng Wang, Yaosen Min, Xin Chen, and Ji Wu. Multi-view graph contrastive representation learning for drug-drug interaction prediction. In *Proceedings of the web conference 2021*, pages 2921–2933, 2021.

[40] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24, 2020.

[41] Feng Xia, Ke Sun, Shuo Yu, Abdul Aziz, Liangtian Wan, Shirui Pan, and Huan Liu. Graph learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2(2):109–127, 2021.

[42] Fan Xu, Nan Wang, Hao Wu, Xuezhi Wen, Dalin Zhang, Siyang Lu, Binyong Li, Wei Gong, Hai Wan, and Xibin Zhao. Gladformer: a mixed perspective for graph-level anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 337–353. Springer, 2024.

[43] Fan Xu, Nan Wang, Hao Wu, Xuezhi Wen, Xibin Zhao, and Hai Wan. Revisiting graph-based fraud detection in sight of heterophily and spectrum. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 9214–9222, 2024.

[44] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[45] Zhenyu Yang, Ge Zhang, Jia Wu, Jian Yang, Shan Xue, Amin Beheshti, Hao Peng, and Quan Z Sheng. Global interpretable graph-level anomaly detection via prototype. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3586–3597, 2025.

[46] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

[47] Ge Zhang, Zhenyu Yang, Jia Wu, Jian Yang, Shan Xue, Hao Peng, Jianlin Su, Chuan Zhou, Quan Z Sheng, Leman Akoglu, et al. Dual-discriminative graph neural network for imbalanced graph-level anomaly detection. *Advances in Neural Information Processing Systems*, 35:24144–24157, 2022.

[48] Lingxiao Zhao and Leman Akoglu. On using classification datasets to evaluate graph outlier detection: Peculiar observations and new insights. *Big Data*, 11(3):151–180, 2023.

# A Preliminary

## A.1 Multi-view Information Bottleneck

The Multi-view Information Bottleneck (MIB) principle [6] offers a rigorous, unsupervised framework for learning representations that distill the essential shared information between multiple views of the same data. It extends the classic Information Bottleneck (IB) [16, 32, 31] to a setting without labels, providing a formal objective for compressing away view-specific noise while retaining a compact summary of the underlying shared semantics.

Given two views, $\mathcal{V}_1$ and $\mathcal{V}_2$, the goal of MIB is to learn a compact latent representation for each view (e.g., $\mathcal{Z}_1$ from $\mathcal{V}_1$) that is maximally predictive of the other view ($\mathcal{V}_2$), while being minimally informative about its own source view ($\mathcal{V}_1$), conditioned on the other view. This formally encourages the model to only retain information that is common to both views. The objective for the latent representation $\mathcal{Z}_1$ is thus formulated as:

$$\max_{\mathcal{Z}_1} \mathcal{I}\big(\mathcal{Z}_1; \mathcal{V}_2\big) - \beta_1 \mathcal{I}\big(\mathcal{Z}_1; \mathcal{V}_1 | \mathcal{V}_2\big), \tag{18}$$

where $\mathcal{I}(\cdot;\cdot)$ denotes the Shannon mutual information (MI). The first term $I(\mathcal{Z}_1; \mathcal{V}_2)$ maximizes the *shared information* by making $\mathcal{Z}_1$ predictive of $\mathcal{V}_2$. The second term $\mathcal{I}(\mathcal{Z}_1; \mathcal{V}_1|\mathcal{V}_2)$ acts as a compression regularizer, penalizing the model for retaining *superfluous information*—that is, any information in $\mathcal{V}_1$ not already present in $\mathcal{V}_2$. The hyperparameter $\beta_1$ balances this trade-off.

Symmetrically, the objective for learning the representation $\mathcal{Z}_2$ from the view $\mathcal{V}_2$ is:

$$\max_{\mathcal{Z}_2} \mathcal{I}\big(\mathcal{Z}_2; \mathcal{V}_1\big) - \beta_2 \mathcal{I}\big(\mathcal{Z}_2; \mathcal{V}_2 | \mathcal{V}_1\big). \tag{19}$$

These objectives provide the theoretical foundation for our approach.

# B Theoretical Analysis and Proofs

## B.1 Derivation of the Unified Information Bottleneck Objective

We derive the unified information bottleneck objective for local information distillation from first principles, using only standard information-theoretic identities and two mild assumptions.

### B.1.1 Definitions

We first define the core concepts and notation used throughout this section.

**Definition 1 (Shannon Entropy and Mutual Information).** For random variables $A, Y, Z$, the Shannon entropy is $\mathcal{H}(X) = -\mathbb{E}_{x\sim\mathcal{P}(X)} \log \mathcal{P}(x)$. The mutual information is $\mathcal{I}(X;Y) = \mathcal{H}(X) - \mathcal{H}(X|Y)$. The conditional mutual information is $\mathcal{I}(X;Y|Z) = \mathcal{H}(X|Z) - \mathcal{H}(X|Y,Z)$.

**Definition 2 (Markov Chain).** Three random variables $X \to Y \to Z$ form a *Markov chain* if and only if:

$$P(Z \mid X, Y) = P(Z \mid Y) \tag{20}$$

or equivalently, $Z \perp\!\!\!\perp X \mid Y$, where $\perp\!\!\!\perp$ denotes statistical independence.

### B.1.2 Causal Model and Core Assumptions

Our pipeline follows a strict causal order:

❖ An original graph $\mathcal{G}$ is sampled from the data distribution $\mathcal{P}(\mathcal{G})$.

❖ Two complementary spectral views are generated from $G$ via deterministic functions:

$$\mathcal{G}^{(lf)} = \mathcal{F}_{lf}(\mathcal{G}), \quad \mathcal{G}^{(hf)} = \mathcal{F}_{hf}(\mathcal{G}) \tag{21}$$

where $\mathcal{F}_{lf}(\cdot)$ and $\mathcal{F}_{hf}(\cdot)$ are the low-pass and high-pass graph filters, respectively.

❖ Core subgraphs are independently extracted from each view via two deterministic subgraph extractor, respectively:

$$\mathcal{G}_s^{(lf)} = f_{lf}^e(\mathcal{G}^{(lf)}), \quad \mathcal{G}_s^{(hf)} = f_{hf}^e(\mathcal{G}^{(hf)}) \tag{22}$$

We make two fundamental assumptions that hold for all standard subgraph extraction methods:

**Assumption 1** (**View Independence**). The extraction process for one view depends only on that view and is conditionally independent of the other view:

$$\mathcal{G}_s^{(lf)} \perp\!\!\!\perp \mathcal{G}^{(hf)} \mid \mathcal{G}^{(lf)}, \quad \mathcal{G}_s^{(hf)} \perp\!\!\!\perp \mathcal{G}^{(lf)} \mid \mathcal{G}^{(hf)} \tag{23}$$

**Assumption 2** (**Subgraph Inclusion**). The extracted subgraph is a subset of the original view, i.e., $\mathcal{G}_s^{(lf)} \subseteq \mathcal{G}^{(lf)}$ and $\mathcal{G}_s^{(hf)} \subseteq \mathcal{G}^{(hf)}$. This implies that the subgraph contains no information beyond what is present in the original view:

$$\mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(lf)}) = 0, \quad \mathcal{H}(\mathcal{G}_s^{(hf)}|\mathcal{G}^{(hf)}) = 0 \tag{24}$$

### B.1.3 Markov Chain Property

We first prove that the causal structure induces a valid Markov chain between the extracted subgraphs.

**Lemma 1** (**Markov Chain Between Subgraphs**). *Under Assumption 1, the following Markov chains hold:*

$$\mathcal{G}_s^{(lf)} \to \mathcal{G}^{(hf)} \to \mathcal{G}_s^{(hf)}, \quad \mathcal{G}_s^{(hf)} \to \mathcal{G}^{(lf)} \to \mathcal{G}_s^{(lf)} \tag{25}$$

*Proof.* We prove the first Markov chain; the second follows by symmetry. By Definition 2, we need to show:

$$P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}_s^{(lf)}, \mathcal{G}^{(hf)}) = P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}^{(hf)}) \tag{26}$$

From Assumption 1, we have:

$$P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}^{(lf)}, \mathcal{G}^{(hf)}) = P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}^{(hf)}) \tag{27}$$

Since $\mathcal{G}_s^{(lf)}$ is a deterministic function of $\mathcal{G}^{(lf)}$, the $\sigma$-algebra generated by $\mathcal{G}_s^{(lf)}$ is a sub-$\sigma$-algebra of that generated by $\mathcal{G}^{(lf)}$. Therefore, conditioning on $\mathcal{G}_s^{(lf)}$ does not provide any additional information beyond conditioning on $\mathcal{G}^{(lf)}$, and we have:

$$P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}_s^{(lf)}, \mathcal{G}^{(hf)}) = P(\mathcal{G}_s^{(hf)} \mid \mathcal{G}^{(lf)}, \mathcal{G}^{(hf)}) \tag{28}$$

Combining the above two equations completes the proof. □

### B.1.4 Original Symmetric Objective

The local information distillation process aims to extract core subgraphs $\mathcal{G}_s^{(lf)}$ (low-frequency view) and $\mathcal{G}_s^{(hf)}$ (high-frequency view) that maximize cross-view shared information while minimizing view-specific redundancy. This is formalized as the symmetric optimization objective:

$$\begin{aligned} \mathcal{L}_{\text{distill}} = & \underbrace{\mathcal{I}(\mathcal{G}^{(lf)}; \mathcal{G}_s^{(lf)} | \mathcal{G}^{(hf)})}_{\text{LF view-specific redundancy}} + \underbrace{\mathcal{I}(\mathcal{G}^{(hf)}; \mathcal{G}_s^{(hf)} | \mathcal{G}^{(lf)})}_{\text{HF view-specific redundancy}} \\ & - \lambda \left[ \underbrace{\mathcal{I}(\mathcal{G}_s^{(lf)}; \mathcal{G}^{(hf)})}_{\text{shared information}} + \underbrace{\mathcal{I}(\mathcal{G}_s^{(hf)}; \mathcal{G}^{(lf)})}_{\text{shared information}} \right] \end{aligned} \tag{29}$$

where $\lambda > 0$ is a trade-off parameter.

### B.1.5 Unified Information Bottleneck Objective

We now derive the unified information bottleneck objective by simplifying Eq. (29).

**Theorem 1** (**Unified IB Objective for Graph Distillation**). *Under Assumptions 1 and 2, the original symmetric objective Eq. (29) can be upper bounded by:*

$$\mathcal{L}_{distill} \leq \mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)}) - \gamma \mathcal{I}(\mathcal{G}_s^{(lf)}; \mathcal{G}_s^{(hf)}) \tag{30}$$

*where $\gamma = 2(1 + \lambda)$. Minimizing this upper bound is sufficient to minimize the original objective.*

*Proof.* We proceed in three steps.

**Step 1: Expand conditional mutual information terms.**

By the chain rule of mutual information (Definition 1):

$$\begin{aligned}\mathcal{I}(\mathcal{G}^{(lf)};\mathcal{G}_s^{(lf)}|\mathcal{G}^{(hf)}) &= \mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(hf)}) - \mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(lf)},\mathcal{G}^{(hf)}) \\ &= \mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(hf)}) - \mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(lf)}) \quad \text{(by Lemma 1)} \\ &= \mathcal{H}(\mathcal{G}_s^{(lf)}|\mathcal{G}^{(hf)}) \quad \text{(by Assumption 2)} \\ &= \mathcal{H}(\mathcal{G}_s^{(lf)}) - \mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)})\end{aligned} \tag{31}$$

By symmetry, we obtain the corresponding result for the high-frequency view:

$$\mathcal{I}(\mathcal{G}^{(hf)};\mathcal{G}_s^{(hf)}|\mathcal{G}^{(lf)}) = \mathcal{H}(\mathcal{G}_s^{(hf)}) - \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)}) \tag{32}$$

**Step 2: Simplify the original objective.**

Substituting these expansions back into the original symmetric objective Eq. (29):

$$\begin{aligned}\mathcal{L}_{\text{distill}} &= \left[\mathcal{H}(\mathcal{G}_s^{(lf)}) - \mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)})\right] + \left[\mathcal{H}(\mathcal{G}_s^{(hf)}) - \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)})\right] \\ &\quad - \lambda\left[\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)}) + \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)})\right] \\ &= \mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)}) - (1+\lambda)\left[\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)}) + \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)})\right]\end{aligned} \tag{33}$$

**Step 3: Lower bound via data processing inequality.**

By the ***data processing inequality***, for any ***Markov chain*** $A \to B \to C$, we have $\mathcal{I}(A;C) \leq \mathcal{I}(B;C)$. From Lemma 1, $\mathcal{G}_s^{(lf)} \to \mathcal{G}^{(hf)} \to \mathcal{G}_s^{(hf)}$ is a valid Markov chain, so:

$$\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}_s^{(hf)}) \leq \mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)}) \tag{34}$$

By symmetry:

$$\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}_s^{(hf)}) \leq \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)}) \tag{35}$$

Adding these two inequalities:

$$\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}^{(hf)}) + \mathcal{I}(\mathcal{G}_s^{(hf)};\mathcal{G}^{(lf)}) \geq 2\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}_s^{(hf)}) \tag{36}$$

Substituting this into the simplified objective gives:

$$\mathcal{L}_{\text{distill}} \leq \mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)}) - 2(1+\lambda)\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}_s^{(hf)}) \tag{37}$$

Letting $\gamma = 2(1+\lambda)$. Since we are minimizing $\mathcal{L}_{\text{distill}}$, minimizing its upper bound is sufficient to ensure the original objective is also minimized. □

That is, we minimize the upper bound:

$$\mathcal{L}_{\text{distill}} = \mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)}) - \gamma\mathcal{I}(\mathcal{G}_s^{(lf)};\mathcal{G}_s^{(hf)}) \tag{38}$$

*Remark* 1. The unified objective Eq. (38) clearly separates the two core requirements of information bottleneck learning:

❖ **Compression**: Minimize the entropy of the extracted subgraphs to ***filter out redundant information***.
❖ **Informativeness**: Maximize the mutual information between the two subgraphs to ***preserve shared task-relevant information***. This objective is general and applies to any deterministic view generation and subgraph extraction process that satisfies Assumptions 1 and 2.

### B.1.6 Implementation Note

The above derivation is performed in the original graph space to establish clear semantic meaning. In practice, since direct computation of mutual information between discrete graph structures is intractable, we follow the standard approach in graph representation learning and approximate the graph-space mutual information $\mathcal{I}(\mathcal{G}_s^{(lf)}; \mathcal{G}_s^{(hf)})$ using the mutual information between their corresponding embeddings $\mathcal{I}(Z^{(lf)}; Z^{(hf)})$.

This approximation is justified by the data processing inequality, which guarantees that the mutual information between embeddings is a lower bound on the mutual information between the original graphs. All theoretical properties derived in the graph space remain valid for the embedding-space approximation. Therefore, substitute it back into unified information bottleneck objective Eq. (38):

$$\mathcal{L}_{\text{distill}} = \underbrace{\mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)})}_{\text{Compression}} - \gamma \underbrace{\mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)})}_{\text{Informativeness}}, \tag{39}$$

The $\mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)})$ term is explicitly optimized using the InfoNCE [27] loss during backpropagation. The entropy term $\mathcal{H}(\mathcal{G}_s^{(lf)}) + \mathcal{H}(\mathcal{G}_s^{(hf)})$ is automatically and implicitly optimized via the SRA mechanism during the forward propagation process, with no need to set an explicit loss term for it or perform backpropagation.

## B.2 SRA as an Implicit Entropy Regularizer

We prove that the ***Score Reconstruction Attention (SRA)*** mechanism inherently minimizes the structural entropy of the distilled subgraph, which exactly implements the *compression term* in our unified information bottleneck objective. Unlike explicit entropy regularization that requires additional loss terms and backpropagation, SRA achieves entropy minimization automatically during the forward propagation process.

### B.2.1 Monotonicity Property of SRA Attention Weights

We first establish the core monotonicity property of SRA, which is the foundation of its entropy regularization effect. Recall the definition of SRA: given raw node importance scores $\boldsymbol{p} = [p_1, p_2, \ldots, p_n]^T \in \mathbb{R}^n$, we define the query, key, and value uniformly: $q_j = p_j$, $k_i = p_i$, $v_i = p_i$. The refined score $p'_j$ for node $j$ is:

$$\begin{aligned} p'_j &= p_j + \sum_{i \in \mathcal{N}(j)} \textit{softmax}_{i \in \mathcal{N}(j)} (q_j k_i)\, v_i \\ &= p_j + \sum_{i \in \mathcal{N}(j)} a_{ji} p_i \end{aligned} \tag{40}$$

where $\mathcal{N}(j)$ denotes the *1-hop neighborhood* of node $v_j$, and $a_{ji} = \textit{softmax}_{i \in \mathcal{N}(j)} (q_j k_i) = \textit{softmax}_{i \in \mathcal{N}(j)} (p_j p_i)$ are the attention weights satisfying $\sum_{i \in \mathcal{N}(j)} a_{ji} = 1$.

In matrix form, SRA can be written as:

$$\boldsymbol{p}' = (\boldsymbol{I} + \boldsymbol{A})\boldsymbol{p} \tag{41}$$

where $\boldsymbol{I}$ is the identity matrix and serves as the residual shortcut connection, and $A$ is the attention weight matrix with entries $\boldsymbol{A}_{ji} = a_{ji}$ if $i \in \mathcal{N}(j)$ and 0 otherwise. By construction, the attention matrix $\boldsymbol{A} = [a_{ji}]$ is *row-stochastic*, i.e., $\sum_{i \in \mathcal{N}(j)} a_{ji} = 1$ for all $j$.

**Lemma 2** (***Monotonicity of SRA***). *For any node $j$ and any two neighbors $i, l \in \mathcal{N}(j)$:*

*If $p_j > 0$, then $p_i > p_l \implies a_{ji} > a_{jl}$. If $p_j < 0$, then $p_i > p_l \implies a_{ji} < a_{jl}$*

*Proof.* Directly follows from the definition of softmax. For $p_j > 0$, $p_i > p_l$ implies $p_j p_i > p_j p_l$, so $\exp(p_j p_i) > \exp(p_j p_l)$, hence $a_{ji} > a_{jl}$. For $p_j < 0$, the inequality direction reverses. □

The Lemma 2 reveals a critical self-reinforcing mechanism of SRA:

❖ High-score nodes ($p_j > 0$) will assign higher attention weights to their high-score neighbors, further boosting their own scores

❖ Low-score nodes ($p_j < 0$) will assign higher attention weights to their low-score neighbors, further suppressing their own scores

### B.2.2 SRA Amplifies Score Differences

We now prove that this self-reinforcing mechanism strictly amplifies the differences between node importance scores.

**Theorem 2** (***Score Difference Amplification***). *For any two nodes $j$ and $k$, if $p_j > p_k$, then $p'_j - p'_k > p_j - p_k$.*

*Proof.* Compute the difference between the refined scores:

$$\begin{aligned} p'_j - p'_k &= \left( p_j + \sum_{i \in \mathcal{N}(j)} a_{ji} p_i \right) - \left( p_k + \sum_{i \in \mathcal{N}(k)} a_{ki} p_i \right) \\ &= (p_j - p_k) + \left( \mathbb{E}_{i \sim a_j(\cdot)}[p_i] - \mathbb{E}_{i \sim a_k(\cdot)}[p_i] \right) \end{aligned} \tag{42}$$

where $\mathbb{E}_{i \sim a_j(\cdot)}[p_i] = \sum_{i \in \mathcal{N}(j)} a_{ji} p_i$ denotes the expected neighbor score under the attention distribution of node $j$.

By Lemma 2, the attention distribution of a higher-score node is always skewed toward higher-score neighbors compared to that of a lower-score node. Therefore:

$$\mathbb{E}_{i \sim a_j(\cdot)}[p_i] > \mathbb{E}_{i \sim a_k(\cdot)}[p_i] \tag{43}$$

Substituting back gives:

$$p'_j - p'_k > p_j - p_k \tag{44}$$

This holds for all pairs of nodes, meaning SRA strictly increases the variance of the node importance scores. □

### B.2.3 Amplified Score Differences Reduce Softmax Entropy

In our implementation, the refined scores $\boldsymbol{p}'$ are passed through a softmax function to obtain the final node retention weights:

$$\tilde{p}_i = \mathit{softmax}(\boldsymbol{p}')_i = \frac{\exp(p'_i)}{\sum_{l=1}^{n} \exp(p'_l)} \tag{45}$$

The structural entropy of the distilled subgraph is defined as the entropy of this categorical distribution:

$$\mathcal{H}(\mathcal{G}_s) = -\sum_{i=1}^{n} \tilde{p}_i \log \tilde{p}_i \tag{46}$$

**Theorem 3** (***Entropy Reduction***). *If SRA amplifies the differences between node scores (Theorem 2), then $\mathcal{H}(\mathit{softmax}(\boldsymbol{p}')) < \mathcal{H}(\mathit{softmax}(\boldsymbol{p}))$.*

*Proof.* For any real vector $\boldsymbol{x}$, the entropy of $\mathit{softmax}(c\boldsymbol{x})$ is a strictly decreasing function of the scaling factor $c > 0$. This is because increasing $c$ amplifies the relative differences between elements of $\boldsymbol{x}$, making the softmax distribution more peaked and thus reducing its entropy.

Theorem 3 shows that SRA performs a transformation on $\boldsymbol{p}$ that strictly increases the relative differences between all pairs of elements. This is equivalent to applying a non-linear transformation that has the same entropy-reducing effect as scaling by $c > 1$. Therefore:

$$\mathcal{H}(\mathit{softmax}(\boldsymbol{p}')) < \mathcal{H}(\mathit{softmax}(\boldsymbol{p})) \tag{47}$$

□

### B.3 Anchor Condensation as an Implicit Entropy Regularizer

We prove that the anchor-guided condensation mechanism inherently minimizes the entropy of the global representation, exactly satisfying the compression requirement of the global information bottleneck objective. Consistent with the local SRA mechanism, this entropy minimization is achieved automatically during forward propagation without any explicit loss terms, additional hyperparameters, or separate backpropagation steps.

#### B.3.1 Global Information Bottleneck Objective

The global information bottleneck objective for cross-view graph distillation is defined as:

$$\mathcal{L}_{\text{global}} = \underbrace{\mathcal{H}(\boldsymbol{a}_G^{(lf)}) + \mathcal{H}(\boldsymbol{a}_G^{(hf)})}_{\text{Global Compression}} - \gamma \underbrace{\mathcal{I}(\boldsymbol{a}_G^{(lf)}; \boldsymbol{a}_G^{(hf)})}_{\text{Global Informativeness}} \tag{48}$$

where $\boldsymbol{a}_G^{(lf)}$ and $\boldsymbol{a}_G^{(hf)}$ are the global anchor representations aggregated from the low-frequency and high-frequency views, respectively. The compression term requires us to minimize the entropy of the global anchor representations to discard redundant information, while the mutual information term ensures that task-relevant information shared between the two views is preserved.

#### B.3.2 Anchor Generation

Given the distilled subgraph representations $\mathbf{Z} \in \mathbb{R}^{n \times d}$, we obtain the compact global anchor representation $\boldsymbol{a}_G \in \mathbb{R}^d$ through a deterministic learnable mapping:

$$\boldsymbol{a}_G = \phi_\theta(\mathbf{Z}) = LastRow(\mathcal{F}_\theta(\mathbf{Z})) \tag{49}$$

where $\phi_\theta(\cdot)$ denotes generation mapping (including feature decoder $\mathcal{F}_\theta$ and row selection). Critically, this mapping is deterministic: given identical node representations $\mathbf{Z}$, it always outputs the same anchor node.

#### B.3.3 Entropy Non-Increase Under Deterministic Transformations

We first state a fundamental result from information theory.

**Lemma 3** (***Deterministic Transformation Lemma***)**.** *For any random variable $\mathbf{X}$ and any deterministic function $f$, the entropy of the transformed variable satisfies:*

$$\mathcal{H}(f(\mathbf{X})) \leq \mathcal{H}(\mathbf{X}) \tag{50}$$

*Equality holds if and only if $f$ is a bijection (one-to-one and onto mapping) that preserves all information in $\mathbf{X}$.*

*Proof.* This is a direct corollary of the data processing inequality. For any deterministic function $f$, the Markov chain $\mathbf{X} \to f(\mathbf{X}) \to$ constant holds, so information can only be lost or preserved, never gained. Equality holds only when no information is lost, i.e., when $f$ is invertible. □

This lemma 3 captures the core intuition of lossy compression: any deterministic transformation that maps a larger space to a smaller space will necessarily *reduce entropy by discarding non-preserved information*.

#### B.3.4 Anchor Condensation Reduces Joint Entropy

We now apply Lemma 3 to our anchor condensation process to formally prove its entropy-reducing property.

**Theorem 4** (***Entropy Reduction via Anchor Node***)**.** *Let $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ be the joint random variable representing all node representations. Then the entropy of the global anchor satisfies:*

$$\mathcal{H}(\boldsymbol{a}_G) < \mathcal{H}(\mathbf{Z}) \tag{51}$$

*for all practical cases with $n > 1$ (multiple nodes).*

*Proof.* By definition, $\boldsymbol{a}_G = \phi_\theta(\mathbf{Z})$ is a deterministic function of the joint node representation $\mathbf{Z}$. Applying Lemma 3 directly gives:

$$\mathcal{H}(\boldsymbol{a}_G) = \mathcal{H}(\phi_\theta(\mathbf{Z})) \leq \mathcal{H}(\mathbf{Z}) \tag{52}$$

For $n > 1$, the functions $\phi_\theta$ maps an $n \times d$-dimensional space to a $d$-dimensional space. This dimensionality reduction means the function cannot be a bijection. Therefore, strict inequality $\mathcal{H}(\boldsymbol{a}_G) < \mathcal{H}(\mathbf{Z})$ always holds. □

Theorem 4 establishes that our anchor generation is inherently an entropy-reducing operation. It compresses the total information of all nodes into a single compact global vector without any explicit regularization.

Combining Theorem 4 with the global information bottleneck objective (Eq. 48), we conclude that anchor condensation automatically implements the global compression term. It reduces the entropy of the global representation by discarding redundant distributed information, while the mutual information term $\mathcal{I}(\boldsymbol{a}_G^{(lf)}; \boldsymbol{a}_G^{(hf)})$ is explicitly optimized using the InfoNCE and ensures that task-relevant cross-view information is preserved.

Notably, this implicit entropy regularization shares exactly the same advantages as the local SRA mechanism: no additional loss terms, no extra hyperparameters, and no separate backpropagation steps. This unifies the local and global compression mechanisms into a single coherent framework, making `Sica` both theoretically elegant and computationally efficient.

### B.4 Anomaly Separation Capacity of Model

#### B.4.1 Theorem

Our theoretical analysis is grounded in the Multi-view Information Bottleneck (MIB) principle. We formalize the intuition that an anomaly is a perturbation that reduces the mutual information between our complementary low- and high-frequency views.

**Theorem 5** (***Anomaly Separation Capacity of the MIB Objective***)**.** *Let $\mathcal{P}_{norm}$ be the distribution of normal graphs. Let $\mathcal{G}_{norm} \sim \mathcal{P}_{norm}$ be a normal graph. Let $\mathcal{G}_{anom} = \mathcal{G}_{norm} + \Delta$ be an anomalous graph created by a perturbation $\Delta$. Let $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)})$ denote the mutual information between the distilled core subgraph derived from a graph $\mathcal{G}$. The ideal anomaly score is $S(\mathcal{G}) = -\mathcal{I}(\mathcal{G})$, and its empirical estimator is the training loss $\mathcal{L}(\mathcal{G})$, such that $\mathbb{E}[\mathcal{L}(\mathcal{G})]$ approximates $S(\mathcal{G})$ up to constants.*

*An anomaly $\Delta$ is defined as MIB-distinguishable if it induces a non-trivial reduction in the core mutual information, such that for some $\epsilon > 0$:*

$$\mathcal{I}(\mathcal{G}_{anom}) \leq \mathcal{I}(\mathcal{G}_{norm}) - \epsilon. \tag{53}$$

*Then, the expected anomaly score computed by `Sica` for the anomalous graph is bounded below by the score for the normal graph plus a margin proportional to the information drop:*

$$\mathbb{E}[\mathcal{L}(\mathcal{G}_{anom})] \geq \mathbb{E}[\mathcal{L}(\mathcal{G}_{norm})] + \epsilon'. \tag{54}$$

*where $\epsilon'$ is a positive value dependent on $\epsilon$ and the tightness of the InfoNCE bound to the true mutual information.*

Theorem 5 provides a theoretical guarantee for our approach. It aligns with our motivation that structural anomalies, such as fraudulent subgroups or spurious connections, will inherently disrupt the shared information between spectral views. Since the InfoNCE loss serves as a tractable estimator for negative mutual information, our scoring function is principled to assign higher scores to these information-theoretic deviations.

#### B.4.2 Proof

*Proof.* Our goal is to prove that the contrastive loss $\mathcal{L}(\mathcal{G})$, acting as a proxy for negative mutual information, provides a formal separation between normal and anomalous graphs. The proof proceeds in three main steps.

**Step 1: InfoNCE as a Variational Lower Bound.** Let $\mathbf{Z}^{(lf)}$ and $\mathbf{Z}^{(hf)}$ be the random variables representing the distilled representations of the low-frequency and high-frequency views, respectively. The InfoNCE loss for a batch size $K$ (one positive pair and $K-1$ negative samples) is defined as:

$$\mathcal{L}(\mathcal{G}) = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(\mathbf{z}_i^{(lf)}, \mathbf{z}_i^{(hf)})/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(\mathbf{z}_i^{(lf)}, \mathbf{z}_j^{(hf)})/\tau)}\right]. \tag{55}$$

Following the derivation by [27], the optimal critic for this contrastive objective satisfies:

$$\mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)}) \geq \log(K) - \mathbb{E}[\mathcal{L}(\mathcal{G})]. \tag{56}$$

Rearranging this inequality, we obtain a lower bound for the expected loss:

$$\mathbb{E}[\mathcal{L}(\mathcal{G})] \geq \log(K) - \mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)}). \tag{57}$$

Let $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\mathbf{Z}^{(lf)}; \mathbf{Z}^{(hf)})$ denote the mutual information between views for a graph $\mathcal{G}$, and let $C = \log(K)$ be a constant determined by the negative sampling strategy. The expected loss can be expressed as:

$$\mathbb{E}[\mathcal{L}(\mathcal{G})] = C - \mathcal{I}(\mathcal{G}) + \delta(\mathcal{G}), \tag{58}$$

where $\delta(\mathcal{G}) \geq 0$ represents the *approximation gap* (or variational gap) resulting from the finite capacity of the encoders and the sub-optimality of the critic.

**Step 2: Information Drop under MIB-Distinguishable Anomalies.** By the definition of an **MIB-distinguishable anomaly**, the perturbation $\Delta$ disrupts the semantic consistency between spectral views. This is formally expressed as a non-trivial reduction in the core mutual information:

$$\mathcal{I}(\mathcal{G}_{\text{anom}}) \leq \mathcal{I}(\mathcal{G}_{\text{norm}}) - \epsilon, \quad \text{for some } \epsilon > 0. \tag{59}$$

This information drop $\epsilon$ quantifies the "semantic dissonance" introduced by the anomaly. In our spectral context, $\Delta$ represents structural aberrations (e.g., fraudulent cliques or spurious long-range bridges) that appear as noise in one spectral band but do not align with the global structure captured by the other.

**Step 3: Derivation of the Separation Margin.** We now compare the expected anomaly scores (training loss) for normal and anomalous graphs. For normal graphs $\mathcal{G}_{\text{norm}}$, let $\delta_{\text{norm}}$ be the approximation gap after convergence:

$$\mathbb{E}[\mathcal{L}(\mathcal{G}_{\text{norm}})] = C - \mathcal{I}(\mathcal{G}_{\text{norm}}) + \delta_{\text{norm}}. \tag{60}$$

For anomalous graphs $\mathcal{G}_{\text{anom}}$, the lower bound from Eq. (58) and the information drop from Eq. (53) yield:

$$\begin{aligned} &\mathbb{E}[\mathcal{L}(\mathcal{G}_{\text{anom}})] = C - \mathcal{I}(\mathcal{G}_{\text{anom}}) + \delta_{\text{anom}} \\ &\geq C - (\mathcal{I}(\mathcal{G}_{\text{norm}}) - \epsilon) + \delta_{\text{anom}} \\ &= (C - \mathcal{I}(\mathcal{G}_{\text{norm}}) + \delta_{\text{norm}}) + \epsilon + (\delta_{\text{anom}} - \delta_{\text{norm}}). \end{aligned} \tag{61}$$

Substituting the expected normal loss from Eq. (60) into Eq. (61):

$$\mathbb{E}[\mathcal{L}(\mathcal{G}_{\text{anom}})] \geq \mathbb{E}[\mathcal{L}(\mathcal{G}_{\text{norm}})] + \epsilon', \tag{62}$$

where we define the separation margin $\epsilon'$ as:

$$\epsilon' = \epsilon + (\delta_{\text{anom}} - \delta_{\text{norm}}). \tag{63}$$

In a well-trained `Sica` model, the encoders are optimized to maximize the mutual information of normal samples, making $\delta_{\text{norm}}$ minimal. For out-of-distribution anomalous samples, the variational approximation $\delta_{\text{anom}}$ is typically larger or equal to $\delta_{\text{norm}}$ due to the mismatch between the anomalous distribution and the learned representation space. Consequently, $\epsilon' \geq \epsilon > 0$. □

The derivation demonstrates that the expected contrastive loss for anomalies is strictly bounded below by the normal loss plus a margin proportional to the information theoretic drop. Thus, the InfoNCE-based scoring function is a principled and theoretically sound estimator for detecting graph-level anomalies.

# C Complexity Analysis and Pseudocode

## C.1 Time Complexity Analysis

We analyze the time complexity of core components in `Sica`:

❖ **Complementary Spectral Views**: Low-frequency GIN encoding takes $\mathcal{O}(L \cdot |\mathcal{E}| \cdot d)$, where $L$ is the number of GIN layers, $|\mathcal{E}|$ denotes the number of edges, and $d$ is the embedding dimension. High-frequency BWGNN encoding takes $\mathcal{O}(C \cdot |\mathcal{E}| \cdot d)$ with $C$ wavelet kernels. The total time for view generation is $\mathcal{O}\big((L + C) \cdot |\mathcal{E}| \cdot d\big)$.

❖ **Soft Subgraph Extraction**: The SRA module and soft distillation take $\mathcal{O}(|\mathcal{V}| \cdot K \cdot d)$, where $|\mathcal{V}|$ is the number of nodes and $K$ is the average node degree.

❖ **Anchor-Guided Extraction**: The shared BWGNN decoder for anchor graphs takes $\mathcal{O}(|\mathcal{E}| \cdot d)$.

❖ **Training & Inference**: The total training time is $\mathcal{O}\big(B \cdot \big[(L+C) \cdot |\mathcal{E}| \cdot d + |\mathcal{V}| \cdot K \cdot d\big]\big)$ ($B$: batch size). Inference avoids backpropagation, with time complexity $\mathcal{O}\big((L + C) \cdot |\mathcal{E}| \cdot d + |\mathcal{V}| \cdot K \cdot d\big)$.

## C.2 Space Complexity Analysis

The space complexity of `Sica` is dominated by feature storage and model parameters: ❖ **Feature Storage**: Batch node embeddings occupy $\mathcal{O}(B \cdot |\mathcal{V}| \cdot d)$. ❖ **Model Parameters**: The total parameters of GIN, BWGNN, SRA, and decoder are $\mathcal{O}(d^2 + C \cdot d^2)$, independent of graph scale. ❖ **Total Space**: $\mathcal{O}(B \cdot |\mathcal{V}| \cdot d + d^2)$, which is *linear space complexity* and scalable to large-scale graphs.

## C.3 Pseudocode

The overall pipeline of `Sica` is summarized in Algorithm 1 and Algorithm 2, including model initialization, training with multi-granularity contrastive learning, and inference for anomaly scoring.

# D Statistics of Datasets

To ensure a reliable and unbiased evaluation, all datasets used in this study are carefully curated following consistent preprocessing and splitting strategies proposed in UB-GOLD [38]. In particular, the normal (ID) and abnormal (OOD) data are strictly separated, and no overlap exists between training and test samples. This guarantees that all models are evaluated under fair and reproducible conditions. Our datasets covers two major categories: **anomaly detection datasets** and **distribution-shift (OOD) detection datasets**. The graph sizes vary across datasets. Molecular datasets consist of relatively small graphs, typically with fewer than 60 nodes. Protein datasets contain graphs of moderate size, ranging from dozens to several hundreds of nodes. Social network datasets exhibit larger graph structures, with REDDIT-BINARY datasets containing several hundred nodes per graph on average. Further statistics details of these datasets are provided in Table 6 and Table 7, respectively.

## D.1 Anomaly Detection Datasets

The first group consists of 8 datasets designed for graph-level anomaly detection, which include both naturally occurring anomalies and class-based ones. These datasets are mainly drawn from the Tox21 challenge [1] and TU benchmark [25], offering diverse structural and semantic characteristics for evaluating different aspects of anomaly detection.

- **Tox21_MMP**: This dataset originates from the Tox21 challenge and focuses on compounds involved in the matrix metalloproteinase (MMP) pathway. Molecules with abnormal MMP activation are treated as anomalies, reflecting intrinsic biochemical irregularities such as those associated with tumor metastasis. This dataset captures genuine biological anomalies rather than label-based outliers.
- **IMDB-BINARY**: A graph classification benchmark constructed from movie collaboration data, each graph corresponds to a film. Nodes denote actors involved in the same movie, and edges indicate co-acting relationships. This dataset is commonly used to study structural patterns of collaboration in the film industry.

**Algorithm 1** Training Pipeline of `Sica`

**Require:** Graph dataset $\mathcal{D} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}, \mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$; Hyperparameters $\alpha, \beta, \gamma, \tau$; GNN layers $L$; Wavelet kernels $C$
**Ensure:** Trained model parameters $\Theta = \{\mathcal{F}_{lf}, \mathcal{F}_{hf}, \mathcal{F}_\theta, \textit{SRA}, \text{Projection}\}$

Initialize model parameters $\Theta$
**for** each batch $\mathcal{B} \in \mathcal{D}$ **do**
  **for** each graph $\mathcal{G} \in \mathcal{B}$ **do**
    $\mathbf{H}^{(lf)} \leftarrow \mathcal{F}_{lf}(\boldsymbol{X})$ {Low-frequency view (e.g., GIN)}
    $\mathbf{H}^{(hf)} \leftarrow \mathcal{F}_{hf}(\boldsymbol{X})$ {High-frequency view (e.g., BWGNN)}
    $\mathbf{h}_G^{(lf)} \leftarrow \text{MeanPool}(\mathbf{H}^{(lf)})$, $\mathbf{h}_G^{(hf)} \leftarrow \text{MeanPool}(\mathbf{H}^{(hf)})$
    $\boldsymbol{p}^{(lf)} \leftarrow \text{Extractor}(\mathbf{H}^{(lf)})$, $\boldsymbol{p}^{(hf)} \leftarrow \text{Extractor}(\mathbf{H}^{(hf)})$
    $\tilde{\boldsymbol{p}}^{(lf)} \leftarrow \textit{SRA}(\boldsymbol{p}^{(lf)})$, $\tilde{\boldsymbol{p}}^{(hf)} \leftarrow \textit{SRA}(\boldsymbol{p}^{(hf)})$
    $\mathbf{K}_s^{(lf)} \leftarrow \text{Broadcast}(\tilde{\boldsymbol{p}}^{(lf)})$, $\mathbf{K}_s^{(hf)} \leftarrow \text{Broadcast}(\tilde{\boldsymbol{p}}^{(hf)})$
    $\mathbf{Z}^{(lf)} \leftarrow \mathbf{H}^{(lf)} \odot \mathbf{K}_s^{(lf)}$, $\mathbf{Z}^{(hf)} \leftarrow \mathbf{H}^{(hf)} \odot \mathbf{K}_s^{(hf)}$
    Construct anchor graph $\tilde{\mathcal{G}} = (\mathcal{V} \cup \{v_a\}, \mathcal{E} \cup \{(v_a, v) | v \in \mathcal{V}\})$
    $\hat{\mathbf{Z}}^{(lf)} \leftarrow F_\theta(\tilde{\mathbf{Z}}^{(lf)})$, $\hat{\mathbf{Z}}^{(hf)} \leftarrow F_\theta(\tilde{\mathbf{Z}}^{(hf)})$
    $\boldsymbol{a}_G^{(lf)} \leftarrow \text{LastRow}(\hat{\mathbf{Z}}^{(lf)})$, $\boldsymbol{a}_G^{(hf)} \leftarrow \text{LastRow}(\hat{\mathbf{Z}}^{(hf)})$
  **end for**
  $\mathcal{L}_{\text{node}} \leftarrow \mathcal{L}^{(n)}_{\text{InfoNCE}}(\mathbf{H}^{(lf)}, \mathbf{H}^{(hf)})$
  $\mathcal{L}_{\text{graph}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\mathbf{h}_G^{(lf)}, \mathbf{h}_G^{(hf)})$
  $\mathcal{L}_{\text{core}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\mathbf{Z}^{(lf)}, \mathbf{Z}^{(hf)})$
  $\mathcal{L}_{\text{anchor}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\boldsymbol{a}_G^{(lf)}, \boldsymbol{a}_G^{(hf)})$
  $\mathcal{L}_{\text{total}} = (\mathcal{L}_{\text{node}} + \mathcal{L}_{\text{graph}}) + \alpha(\mathcal{L}_{\text{core}} + \mathcal{L}_{\text{anchor}})$
  Back-propagate $\mathcal{L}_{\text{total}}$ to update $\Theta$
**end for**
**return** $\Theta$

**Algorithm 2** Inference Pipeline of `Sica`

**Require:** Test graph $\mathcal{G}$; Trained parameters $\Theta$; Hyperparameter $\beta$
**Ensure:** Anomaly score $s(\mathcal{G})$ (higher = more anomalous)

Compute $\mathbf{H}^{(lf)}, \mathbf{H}^{(hf)}, \mathbf{h}_G^{(lf)}, \mathbf{h}_G^{(hf)}, \mathbf{Z}^{(lf)}, \mathbf{Z}^{(hf)}, \boldsymbol{a}_G^{(lf)}, \boldsymbol{a}_G^{(hf)}$ (Line 4-13 in Alg.1)
$s_{\text{node}} \leftarrow \mathcal{L}^{(n)}_{\text{InfoNCE}}(\mathbf{H}^{(lf)}, \mathbf{H}^{(hf)})$
$s_{\text{graph}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\mathbf{h}_G^{(lf)}, \mathbf{h}_G^{(hf)})$
$s_{\text{core}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\mathbf{Z}^{(lf)}, \mathbf{Z}^{(hf)})$
$s_{\text{anchor}} \leftarrow \mathcal{L}^{(g)}_{\text{InfoNCE}}(\boldsymbol{a}_G^{(lf)}, \boldsymbol{a}_G^{(hf)})$
$s(\mathcal{G}) = s_{\text{node}} + s_{\text{graph}} + \beta(s_{\text{core}} + s_{\text{anchor}})$
**return** $s(\mathcal{G})$

- **REDDIT-BINARY**: A collection of graphs derived from Reddit discussion threads. In each graph, nodes represent users participating in a thread, while edges capture interactions between users. The dataset reflects the structural characteristics of online discussions and their interaction patterns.
- **ENZYMES**: In ENZYMES, each graph represents a protein tertiary structure, with nodes denoting secondary structural elements and edges indicating spatial adjacency. Six enzyme classes are included, and rare classes are regarded as anomalies to evaluate models' sensitivity to functional diversity.
- **PROTEINS**: In PROTEINS, each graph describes complete protein molecule, where nodes correspond to amino acids and edges reflect physical or functional interactions. Anomalous samples are defined as those belonging to minority classes or exhibiting unusual structural organization.
- **DD**: DD comprises protein–protein interaction (PPI) networks, where nodes are individual proteins and edges capture their physical or functional connections. The graphs vary significantly in size and density, making DD suitable for studying structural deviations within biological systems.

Table 6: Statistical details of GLAD datasets.

| Dataset | Domain | # Normal Train | # Normal Test | # Abnormal Test | Anomaly Ratio (%) |
|---|---|---|---|---|---|
| Tox21_MMP | Molecules | 6170 | 200 | 38 | 0.59 |
| IMDB-BINARY | Social Networks | 400 | 100 | 100 | 16.67 |
| REDDIT-BINARY | Social Networks | 800 | 200 | 200 | 16.67 |
| ENZYMES | Proteins | 400 | 100 | 20 | 3.85 |
| PROTEINS | Proteins | 360 | 90 | 133 | 22.81 |
| DD | Proteins | 390 | 97 | 139 | 22.20 |
| BZR | Molecules | 69 | 17 | 64 | 42.67 |
| AIDS | Molecules | 1280 | 320 | 80 | 4.76 |
| COX2 | Molecules | 81 | 21 | 73 | 41.71 |
| DHFR | Molecules | 368 | 93 | 59 | 11.35 |

Table 7: Statistical details and abbreviations of GLOD dataset pairs. "A $\rightarrow$ B" denotes that graphs from A are treated as ID, while graphs from B are treated as OOD.

| ID Dataset | OOD Dataset | Abbreviation | Domain | # ID Train | # ID Test | # OOD Test | OOD Ratio (%) |
|---|---|---|---|---|---|---|---|
| IMDB-MULTI | IMDB-BINARY | IM$\rightarrow$ IB | Social Networks | 1350 | 150 | 150 | 9.09 |
| ENZYMES | PROTEINS | EN$\rightarrow$PR | Proteins | 540 | 60 | 60 | 9.09 |
| AIDS | DHFR | AI$\rightarrow$DH | Molecules | 1800 | 200 | 200 | 9.09 |
| BZR | COX2 | BZ$\rightarrow$CO | Molecules | 364 | 41 | 41 | 9.19 |
| BBBP | BACE | BB$\rightarrow$BA | Molecules | 1835 | 204 | 204 | 9.09 |
| PTC_MR | MUTAG | PT$\rightarrow$MU | Molecules | 309 | 35 | 35 | 9.23 |

- **BZR**: BZR contains benzodiazepine receptor ligands, represented as molecular graphs with atoms as nodes and bonds as edges. The anomalous graphs correspond to molecules exhibiting uncommon binding behaviors toward the receptor.
- **AIDS**: Each graph represents a chemical compound tested for anti-HIV activity. The dataset is often used in antiviral drug discovery. Here, minority-class compounds are regarded as anomalous, reflecting rare or atypical structural motifs in the molecular space.
- **COX2**: This dataset consists of cyclooxygenase-2 inhibitors used to study anti-inflammatory drug properties. Structural diversity within the compound pool provides a meaningful basis for defining rare or outlier molecules.
- **DHFR**: DHFR includes molecules that inhibit dihydrofolate reductase, an enzyme critical in cancer and bacterial infection treatment. The dataset contains chemically diverse inhibitors, and the rare structural types are considered anomalous.

### D.2 Out-of-distribution Detection Datasets

The second group focuses on inter-dataset distribution shift, which simulates real-world OOD detection scenarios. Following the setup in UB-GOLD, each OOD task is constructed by pairing two semantically related datasets that differ in domain or data distribution. The first dataset in each pair serves as the ID domain, while the second acts as OOD. This setup evaluates how well a model generalizes beyond the training distribution. The abbreviations of these dataset pairs are summarized in Table 7.

- **IMDB-MULTI$\rightarrow$IMDB-BINARY**: Both datasets describe social networks built from IMDB movie collaborations, where nodes represent actors and edges represent their co-starring in movies. IMDB-MULTI includes multiple genres, whereas IMDB-BINARY merges them into two classes. Despite sharing the same data source, the label simplification and sampling distribution introduce a clear shift in graph characteristics.
- **ENZYMES$\rightarrow$PROTEINS**: This pair combines two protein-related datasets with different levels of biological abstraction. ENZYMES focuses on tertiary structures, while PROTEINS captures amino acid–level interactions. The structural differences between them form a meaningful inter-domain gap.
- **AIDS$\rightarrow$DHFR**: Both datasets contain molecular graphs but originate from distinct biochemical contexts—anti-HIV activity versus enzyme inhibition. Although their graph structures share similar chemical motifs, their underlying functional properties are fundamentally different.

- **BZR→COX2**: These two molecular datasets describe ligands targeting different receptors. BZR molecules bind to benzodiazepine receptors, whereas COX2 molecules inhibit cyclooxygenase-2. The task examines whether a model trained on one biochemical mechanism can identify OOD samples from another.
- **BBBP→BACE**: BBBP contains data on Blood–brain barrier permeability (BBBP), while BACE provides $\beta$-secretase inhibitor information. Both relate to drug discovery but differ in biological targets and labeling schemes, providing a realistic domain shift scenario.
- **PTC_MR→MUTAG**: Both datasets belong to the toxicology domain. PTC-MR consists of compounds tested for carcinogenicity in male rats, and MUTAG contains mutagenicity-labeled molecules. The biological relevance is related, but the statistical distributions differ significantly, producing a challenging OOD setup.

# E Details of Baselines

To comprehensively evaluate our proposed framework, we compare it against a diverse set of baseline methods that cover both classical and graph neural network–based approaches. Following the taxonomy introduced in UB-GOLD, these methods can be grouped into 2 categories: GNN-based graph-level anomaly detection (GLAD) models, and graph OOD detection approaches. All baselines are implemented using the official or widely adopted public code, with hyperparameters tuned to achieve their reported performance.

## E.1 GNN-Based End-to-End GLAD Methods

Recent graph-level anomaly detection approaches leverage graph neural networks to learn expressive representations in an end-to-end manner. These methods often integrate self-supervised or one-class objectives to identify anomalous graphs based on structural or semantic deviations.

- **OCGIN** [48]: OCGIN employs a Graph Isomorphism Network (GIN) encoder trained under a Support Vector Data Description (SVDD) objective. By minimizing the distance of normal samples to a learned hypersphere center, OCGIN effectively detects anomalous graphs while maintaining interpretability.
- **OCGTL** [28]: OCGTL extends deep one-class classification to graphs through transformation learning. It introduces graph transformation regularization, which enhances the model's capacity to detect structural anomalies without supervision.
- **SIGNET** [19]: SIGNET presents a self-interpretable framework that maximizes mutual information among subgraphs from multiple views. It jointly performs anomaly detection and subgraph-level interpretation, offering insights into why a graph is identified as abnormal.
- **CVTGAD** [13]: CVTGAD integrates a lightweight transformer with cross-view attention for unsupervised graph anomaly detection. By capturing both intra- and inter-graph dependencies, it compensates for the limited receptive field of traditional GNNs and improves detection robustness.
- **MUSE** [11]: MUSE is a graph autoencoder-based method for graph-level anomaly detection. It utilizes multifaceted summaries of reconstruction errors as graph features, surpassing the mean reconstruction error and capturing richer structural and attribute deviations.
- **GLADPro** [45]: GLADPro is a prototype-based interpretable GLAD model with the information bottleneck principle. It learns global explanatory subgraph prototypes, applies regularization and prototype filtering to avoid collapse, and achieves high detection accuracy with global explainability for graph anomalies.

## E.2 GNN-Based End-to-End GLOD Methods

In addition to anomaly detection, we benchmark against representative methods for graph-level out-of-distribution (OOD) detection, which focus on identifying graphs sampled from unseen data distributions. These approaches differ in whether they require model retraining or operate at test time only.

- **GOOD-D** [18]: GOOD-D adopts a perturbation-free contrastive learning framework with hierarchical consistency objectives. By contrasting graphs at multiple levels, it learns distributionally robust representations that generalize well to unseen domains.
- **SEGO** [9]: SEGO is an unsupervised contrastive learning framework for graph out-of-distribution detection. It leverages structural entropy to build coding tree anchor views for redundant structure removal, and adopts multi-grained contrastive learning to effectively distinguish ID and OOD graphs with stable detection performance.
- **GOODAT** [36]:A test-time OOD detection method designed to work with pretrained GNNs. It incorporates a graph masker and the Graph Information Bottleneck principle to extract the most informative subgraphs. The GIB-guided loss enhances sensitivity to distributional shifts, enabling strong detection performance across heterogeneous datasets.

# F More Experiments

## F.1 Full Component Ablation

Table 8 reports the complete component ablation results on all 16 datasets, merging the four main benchmarks and the additional 8 GLAD and 4 GLOD datasets into a single table. The full model achieves the best average AUROC of 81.63%, consistently outperforming all ablated variants. These results are fully consistent with the conclusions drawn in the main text: the anchor-guided condensation is the most critical component, the band-pass decoder and spectral views provide essential complementary graph signals, and SRA further encourages the extraction of cohesive substructures. We note a few dataset-level fluctuations. For example, on REDDIT-B, removing SRA yields 86.28% versus 85.77% of the full model, and on DD, removing the anchor node yields 79.02% versus 79.12% of the full model. Such minor fluctuations do not change the overall conclusion, the full model remains the best on average.

Table 8: Full model component ablation study of `Sica` on all 16 datasets (AUROC %, mean). Each component is removed to verify its contribution.

| Dataset | w/o spectral views | w/o band-pass decoder | w/o SRA | w/o anchor node | Full |
|---|---|---|---|---|---|
| BZR | 79.12 | 73.25 | 80.44 | 70.24 | **84.27** |
| COX2 | 72.21 | 67.58 | 72.67 | 65.45 | **74.83** |
| DHFR | 68.11 | 67.28 | 70.28 | 63.67 | **71.34** |
| AIDS | 96.07 | 99.23 | 94.68 | 92.54 | **99.54** |
| IMDB-B | 61.69 | 60.43 | 64.52 | 56.81 | **65.73** |
| REDDIT-B | 81.93 | 84.04 | **86.28** | 82.95 | 85.77 |
| ENZYMES | 65.07 | 57.42 | 62.32 | 60.66 | **66.44** |
| PROTEINS | 77.14 | 77.75 | 77.85 | 75.93 | **78.26** |
| DD | 76.56 | 76.47 | 77.82 | 79.02 | **79.12** |
| MMP | 70.54 | 69.30 | 69.39 | 65.66 | **71.41** |
| PT→MU | 87.28 | 78.82 | 91.57 | 77.90 | **92.11** |
| EN→PR | 64.03 | 64.24 | 64.07 | 57.98 | **68.24** |
| AI→DH | 99.43 | 97.63 | 99.55 | 96.10 | **99.91** |
| BZ→CO | 98.79 | 95.86 | 97.43 | 90.98 | **99.31** |
| BB→BA | 84.31 | 83.87 | 84.14 | 79.76 | **85.53** |
| IM→IB | 76.85 | 82.39 | 78.82 | 74.37 | **84.27** |
| Avg. AUROC | 78.70 | 77.22 | 79.49 | 74.38 | **81.63** |

## F.2 Efficiency Evaluation

We evaluate the runtime efficiency of state-of-the-art methods on the DHFR dataset. During evaluation, we measure inference latency and throughput. All experiments are conducted on a single NVIDIA GeForce RTX 4060 GPU with a batch size of 128. As shown in Table 9, `Sica` maintains a moderate number of parameters and achieves competitive computational efficiency, indicating a favorable trade-off between model capacity and efficiency compared to other baselines.

Table 9: Efficiency evaluation on DHFR datasets.

| Model | # Params (M) | Inference Latency (ms) | Throughput (samples/s) | Throughput(batches/s) |
|---|---|---|---|---|
| OCGTL | 0.134 | 51.970 | 2263.493 | 17.684 |
| OCGIN | 0.023 | 13.565 | 7084.936 | 55.351 |
| SIGNET | 0.121 | 12.828 | 7564.692 | 59.099 |
| CVTGAD | 9.715 | 27.417 | 4447.310 | 34.745 |
| MUSE | 1.148 | 2.079 | 17904.052 | 139.875 |
| GOOD-D | 0.311 | 6.355 | 13246.479 | 103.488 |
| `Sica` | 1.408 | 14.421 | 6706.280 | 52.393 |

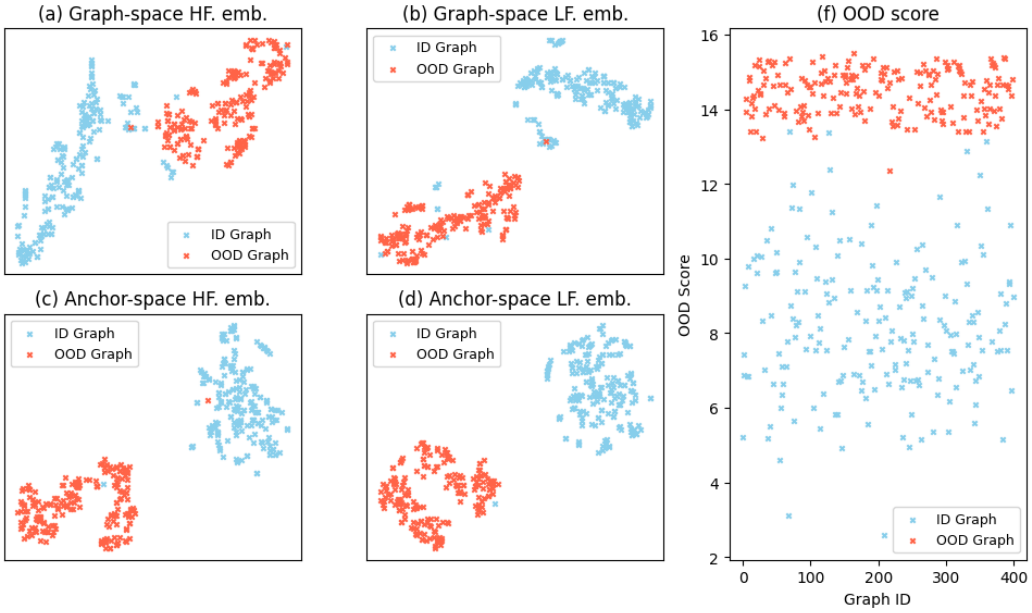



Figure 5: Visualization of learned graph representations on AIDS→DHFR dataset pair.

### F.3 Visualization

To provide an intuitive understanding of the representations learned by `Sica`, we visualize graph-level embeddings on the representative dataset pair **AIDS→DHFR**. Figure. 5 shows **t-SNE** [24] projections from complementary spectral views, including the graph space ($\mathbf{h}_G^{(lf)}$, $\mathbf{h}_G^{(hf)}$) and the condensed anchor space ($\boldsymbol{a}_G^{(lf)}$, $\boldsymbol{a}_G^{(hf)}$).

As shown in Figure. 5(a) and 5(b), the full-dataset embeddings from the low-frequency and high-frequency spectral views already exhibit clear separation between ID (AIDS) and OOD (DHFR) graphs. This validates that our semantically complementary design effectively captures the global structural differences between distributions, rather than just local node features. However, the embedding clusters are relatively diffuse across the full dataset, meaning redundant structural information is still preserved in the raw spectral views.

In contrast, Figure. 5(c) and 5(d) show that anchor-space embeddings form more compact and separated clusters. This trend holds uniformly across all samples: intra-class variance is reduced, inter-class distance is enlarged, and two spectral views are aligned into a unified global space. This observation indicates that `Sica` suppresses superfluous spectral components and aligns spectral signals to the anchor node. This property is particularly meaningful for ***molecular graphs***. In AIDS→DHFR, ID and OOD graphs share similar local substructures but differ in global semantics. By aligning distilled representations across complementary spectral views, the anchor node captures global distributional characteristics, enabling `Sica` to distinguish graphs that are locally similar but globally shifted.

Furthermore, (f) visualizes the OOD scores generated by `Sica`, which exhibit a clear margin between the ID and OOD graphs. This aligns with the observed embedding separation phenomenon, confirming that these learned embeddings can be directly translated into unbiased anomaly scores or OOD scores.

## G Further Discussion: Advantages and Limitations

**Performance about GLAD.** In the GLAD task, from the AUROC performance perspective, `Sica` attains substantial and consistent gains on molecular and protein graphs (e.g., BZR, COX2, PROTEINS, DD) while yielding relatively moderate improvements on social network datasets (IMDB-B, REDDIT-B). Such a discrepancy arises from the inherent structural differences across domains. Molecular and protein graphs possess well-defined, functionally critical substructures, where anomalies typically present as either low-frequency global aberrations (e.g., skeleton distortion, deviant community layouts) or high-frequency local heterophilous variations (e.g., unusual chemical bonds, sparse long-range connections). The two complementary spectral views in `Sica` explicitly cover these two categories of patterns: the low-frequency view captures stable global homophily and community structure, and the high-frequency view retains sensitive local variations that low-pass GNNs tend to smooth out. Coupled with the Score Reconstruction Attention (SRA) and anchor-guided condensation, the framework distills compact, task-relevant substructures and filters view-specific noise, rendering high sensitivity to subtle yet crucial structural anomalies. In contrast, social network graphs are larger, noisier, and lack clearly defined structural anomalies; anomalies are often defined by label imbalance rather than topological irregularities, where high-frequency signals are weak and easily overwhelmed by background noise. As a result, the gain brought by complementary spectral views is relatively ***constrained*** in such scenarios.

**Performance about GLOD.** In the GLOD task, however, `Sica` outperforms all baselines by a clear margin across all dataset pairs, including social network distribution shifts (e.g., IMDB-MULTI→IMDB-BINARY). This distinct performance reveals the strong distribution generalization ability of `Sica`, which ***stems from the mismatch between GLAD and GLOD in learning objectives***. GLAD focuses on detecting structural deviations within a single dataset, while GLOD aims to identify global distributional differences across datasets. The multi-view information bottleneck principle in `Sica` compresses redundant view-specific information and retains shared core semantics across complementary spectral views, and the anchor-guided condensation further aggregates global structural patterns into a unified compact embedding. Such a design is inherently adept at capturing cross-dataset distribution discrepancies rather than mere local anomalies. For social network GLOD tasks, the distribution shift mainly comes from sampling strategies and global community connectivity patterns rather than local node-level abnormalities. The global alignment and information compression mechanisms of `Sica` precisely match this characteristic, enabling reliable discrimination between in-distribution and out-of-distribution samples even on noisy social network graphs.

**Mechanistic Level.** the efficacy of `Sica` relies on three synergistic designs. First, the semantically complementary spectral views replace redundant homogeneous augmentations, covering the full spectrum of anomalous patterns from global low-frequency structures to local high-frequency variations, which resolves the issue of anomaly masking in traditional methods. Second, the soft subgraph extractor with SRA implicitly minimizes the structural entropy of distilled subgraphs, and the anchor-guided condensation further reduces the entropy of global representations, forming a **dual information bottleneck that filters noise and preserves essential shared information**. Third, the contrastive objective enforces hierarchical alignment from nodes to anchors, which amplifies subtle structural deviations while maintaining the spectral properties of original views. The strong robustness against feature noise further validates the stability of the information bottleneck and soft distillation mechanisms, making `Sica` adaptable to real-world noisy graph data.

**The Advantages of Model.** `Sica` is optimally suited for molecular and protein graphs with clear structural semantics and well-defined anomalies, where it delivers state-of-the-art GLAD and GLOD performance. For noisy social network graphs, though its advantage is milder in anomaly detection, it excels in out-of-distribution detection with strong generalization. By unifying the design for both GLAD and GLOD tasks, `Sica` provides a generic solution for unsupervised graph-level anomaly and out-of-distribution detection.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the core technical contributions, research scope and inherent assumptions of the proposed method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: A discussion section in appendix discusses model generalization boundaries, dataset limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical theorems are equipped with explicit assumptions, with complete formal proofs provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Dataset, model architectures and experimental implement are detailed to replicate all experimental conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

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

Answer: [Yes]

Justification: Anonymized code and data preprocessing scripts are released in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All hyperparameters configurations and data strategies are documented.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Standard deviations across multiple runs are reported to verify statistical significance of results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provides information on the computer resources in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The societal impacts of this work are discussed in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmark datasets and baseline implementations are properly cited with clear license information and original references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: No new assets are released as part of this paper submission.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: This research involves no crowdsourcing tasks or any experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research is conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not involved in core algorithm design or methodological innovation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.