PEACE: PERTURBATION–PROPAGATION EVIDENCE CERTIFICATION WITH GEOMETRIC ELIMINATION FOR POSITIVE-UNLABELED GRAPH CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the problem of positive-unlabeled (PU) graph classification, which aims to learn graph-level classifiers from labeled positive graphs and unlabeled graphs with no confirmed negative graphs. Existing approaches usually incorporate pseudo-labeling techniques into graph neural networks (GNNs). Despite recent progress, the performance of these approaches is still far from satisfactory due to the lack of negative evidence and thus noisy pseudo-labels. Towards this end, we propose a novel approach named Perturbation–Propagation Evidence Certification with Geometric Elimination (**PEACE**) for PU graph classification. The core of our **PEACE** is to identify reliable supervision signals in unlabeled graphs from two complementary sources of evidence, *i.e.*, perturbation consistency and relational propagation. On the one hand, our **PEACE** constructs multi-granularity subgroups of augmented graph views through random perturbation to assess perturbation consistency of pseudo-labels for unlabeled graphs guided by Pareto dominance. On the other hand, **PEACE** constructs a graph-of-graphs over graph samples to capture inter-graph dependencies and utilize label propagation to derive relational evidence. The two complementary sources of evidence are then integrated from a distribution view to produce reliable pseudo-labels for unlabeled graphs. Furthermore, we exploit the geometric structure of the latent space to estimate prototypical deviance, which dynamically guides the elimination of noisy graph samples for smooth optimization. Extensive experiments on multiple graph classification benchmarks demonstrate that **PEACE** consistently outperforms state-of-the-art PU graph classification approaches.

1 INTRODUCTION

Positive–unlabeled (PU) graph classification aims to learn a graph-level classifier from a small set of labeled positive graphs and a large set of unlabeled graphs containing both positives and negatives (Zhao et al., 2011). It has been widely applied to real-world tasks such as molecular screening (Gong et al., 2026), biological function prediction (Zhapa-Camacho et al., 2024), and fraud detection (Sun et al., 2026), where obtaining reliable graph-level annotations for all samples is often costly and labor-intensive. The fundamental challenge is that positive and negative graphs are difficult to distinguish within the unlabeled set when only limited positive supervision is available.

Existing PU learning methods for graph-structured data can be broadly divided into two-step methods (Liang et al., 2024) and end-to-end methods (Liang et al., 2025). The former typically identify reliable negative graphs from the unlabeled set and then train a binary model together with the labeled positives based on graph features, structural patterns, or graph similarities (Zhao et al., 2011; Liang et al., 2024). However, these methods often struggle when complex graph structures make reliable negative selection difficult. The latter directly learn graph representations from positive and unlabeled graphs by incorporating risk estimation, sample weighting, or other learning objectives to reduce the influence of unlabeled samples (Kumagai et al., 2025; Liang et al., 2025). Nevertheless, most existing methods still rely on supervision inferred from unlabeled graphs during training, whose reliability may not always be guaranteed in real-world scenarios.

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

In this paper, we study reliable and self-correcting pseudo-labeling for PU graph classification. Achieving this goal is non-trivial, as it entails addressing two interrelated challenges. ❶ ***How can reliable information be obtained from unlabeled graphs under limited supervision?*** In PU graph classification, the unlabeled set contains both positive and negative graphs, while only a small subset of positives is labeled. Under such limited supervision, model predictions may not provide sufficient evidence to reliably assess unlabeled graphs (Kim et al., 2026). Consequently, incorrect pseudo-labels may be mistakenly introduced as supervision and impair subsequent learning. ❷ ***How can the adverse effects of unreliable supervision be mitigated during training?*** In practice, pseudo-labels generated at an early stage may become unreliable as training proceeds. However, these labels are often treated as fixed supervision once assigned, without being reconsidered during subsequent optimization. Such persistent errors can be repeatedly reinforced during training, further amplifying confirmation bias and degrading classification performance.

To address the above challenges, we propose Perturbation–Propagation Evidence Certification with Geometric Elimination (named **PEACE**) for PU graph classification. The goal of **PEACE** is to improve pseudo-label reliability while preventing erroneous supervision from being reinforced during training. Given that a single prediction confidence is insufficient to characterize the uncertainty of an unlabeled graph, **PEACE** evaluates each graph from two complementary views. It generates multiple perturbed graph views and adaptively aggregates their predictions to capture intra-graph consistency, while constructing a graph-of-graphs structure to exploit inter-graph relations among structurally similar graphs. The two types of evidence are integrated to obtain more reliable pseudo-labels. Moreover, even initially confident pseudo-labels may become inconsistent with their assigned classes as training proceeds. To mitigate this issue, **PEACE** re-evaluates selected pseudo-labeled graphs according to their distances from the corresponding class prototypes and filters suspicious assignments from subsequent supervision. In this way, **PEACE** improves pseudo-label reliability while reducing the accumulation of labeling errors during training. We summarize our contributions as follows:

❶ ***New Perspective***. We rethink PU graph classification from a complementary and dynamic reliability perspective, extending pseudo-label reliability assessment beyond prediction confidence to multiple graph perspectives and subsequent model learning.

❷ ***Novel Methodology***. We propose **PEACE**, which jointly considers intra-graph structural consistency and inter-graph distributional information for reliable pseudo-labeling, followed by prototype-guided dynamic noise elimination to suppress unreliable supervision.

❸ ***Empirical Validation***. Extensive experiments on graph classification benchmarks under different observed-positive ratios demonstrate the superiority of **PEACE** over PU learning baselines, while ablation and sensitivity studies verify its effectiveness and robustness.

2 Related Work

Positive-Unlabeled Graph Classification. PU graph classification has been studied to learn graph-level prediction models from limited labeled positive graphs and mixed unlabeled graphs, with representative approaches including reliable negative identification (Wu et al., 2016), multi-view graph learning (Liu et al., 2025), and reweighting strategies (Kim et al., 2026). Early methods jointly identify reliable negative graphs and discriminative subgraph features, since informative graph patterns and negative samples are mutually dependent under PU supervision (Wu et al., 2016). Subsequent studies exploit complementary graph views and progressive learning strategies to better characterize unlabeled graphs and reduce the influence of uncertain samples (Liu et al., 2023; 2025). More recently, GNN-based methods incorporate adaptive reweighting into representation learning, emphasizing informative substructures while suppressing ambiguous unlabeled graphs (Kim et al., 2026). Despite these advances, the reliability of unlabeled graphs is largely handled through sample selection or weighting during learning, while complementary evidence for reliability assessment and subsequent correction of erroneous supervision remain less explored.

Semi-Supervised Graph Learning. Semi-supervised graph learning has been extensively studied to exploit abundant unlabeled data when graph annotations are scarce (Ju et al., 2024a). Existing approaches generally enhance supervision through pseudo-labeling (Koh et al., 2024), consistency learning (Wang et al., 2024), or structure-aware representation learning (Ju et al., 2024b). Pseudo-labeling methods expand the labeled set using confident model predictions, whereas consistency-

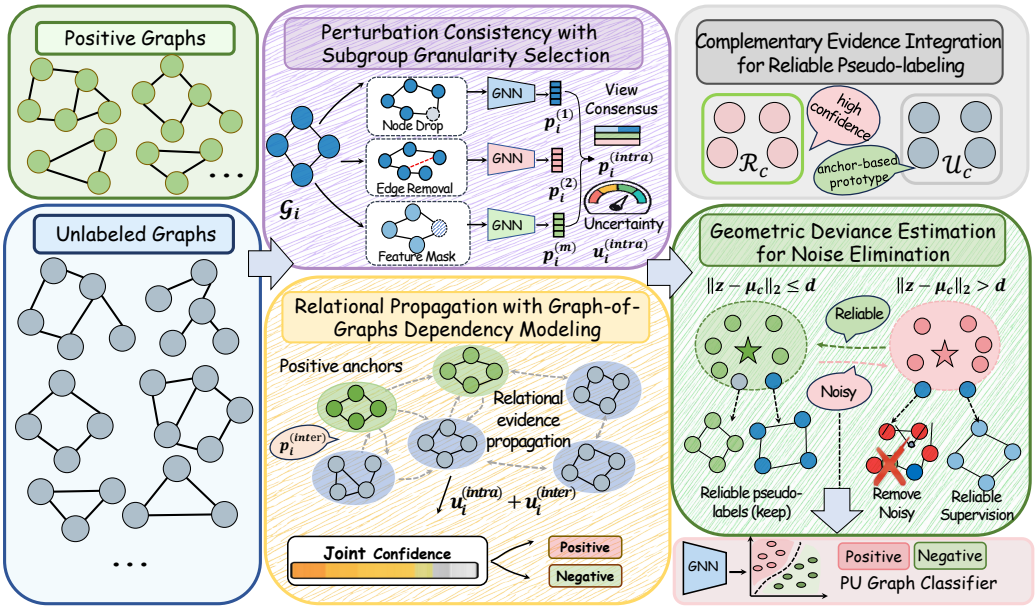


Figure 1: The overall framework of the proposed PEACE. PEACE identifies reliable pseudo-labels by combining perturbation consistency with relational propagation. The resulting supervision is further refined by latent-space geometry to eliminate noisy samples during optimization.

based methods encourage stable predictions or representations across different graph views, models, or perturbations. Another strand of research further exploits structural relations among graphs or complementary graph encoders to transfer information from labeled to unlabeled data (Ju et al., 2023). These strategies demonstrate that unlabeled graphs can provide useful structural and relational information beyond direct label supervision. However, standard semi-supervised graph classification generally assumes that labeled examples from the target classes are available, whereas PU graph classification provides verified labels only for the positive class and leaves the class membership of all remaining graphs unknown. Therefore, our work focuses on reliability estimation under this asymmetric supervision setting, where unlabeled graphs must be carefully assessed before being incorporated into subsequent training. Despite these advances, supervision derived from unlabeled graphs can still be unreliable and introduce errors into subsequent learning. Therefore, our study focuses on improving the reliability of unlabeled information throughout PU graph learning.

3 The Proposed PEACE

Problem Definition. We consider the task of PU graph classification on attributed graphs, where only verified positive graphs are provided during training and the unlabeled set contains an unknown mixture of positive and negative graphs. Let $\mathcal{D}$ be a graph-level PU training set containing N graphs. The i-th graph is $G_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i)$, where $\mathcal{V}_i$ and $\mathcal{E}_i$ are its node and edge sets, and $\mathbf{X}_i \in \mathbb{R}^{|\mathcal{V}_i| \times d}$ is its node-feature matrix with feature dimension d. The training set consists of a positive set $\mathcal{P}$ and an unlabeled set $\mathcal{U}$, where graphs in $\mathcal{P}$ have known positive labels $y_i = 1$, while the ground-truth labels of graphs in $\mathcal{U}$ are unknown. Given only $\mathcal{P}$ and $\mathcal{U}$, the goal is to learn a graph classifier f_θ that reliably distinguishes positive and negative graphs in $\mathcal{U}$ without access to verified negative labels. This raises two related questions: ***(1) How can the reliability of unlabeled graphs be effectively assessed?*** and ***(2) How can the influence of unreliable pseudo-labels be reduced during training?***

3.1 Framework Overview

To address these issues, we propose PEACE, which improves pseudo-label reliability through complementary graph evidence and dynamic noise elimination. First, perturbation consistency with subgroup granularity selection aims to obtain reliable information from individual unlabeled graphs by assessing prediction stability across multiple perturbed views. Next, relational propagation with graph-of-graphs dependency modeling captures dependencies among training graphs and propa-

gates class evidence to obtain relational estimates. Then, complementary evidence integration for reliable pseudo-labeling combines the two forms of evidence to identify reliable pseudo-labels for subsequent optimization. Finally, geometric deviance estimation for noise elimination focuses on reducing the influence of erroneous pseudo-labels during subsequent training by re-examining their consistency with the learned class structure. The overall framework of **PEACE** is illustrated in Figure 1, with the detailed formulations presented in the following subsections.

3.2 Perturbation Consistency with Subgroup Granularity Selection

In PU graph classification, model predictions on unlabeled graphs provide the primary basis for constructing pseudo-label supervision. However, prediction confidence alone may not reliably reflect label quality because predictions can be sensitive to local structural or feature perturbations (Liu et al., 2026). Accordingly, inspired by subgroup-based consensus estimation, we generate multiple perturbed views for each unlabeled graph and evaluate their prediction consistency under different subgroup granularities to obtain reliable pseudo-label candidates.

For each graph G_i, we generate M augmented views $G_i^{(m)} = \mathcal{T}_\rho^{(m)}(G_i)$, where $m \in \{1, \dots, M\}$ indexes the augmented views, $\rho \in [0, 1]$ controls the perturbation ratio, and $\mathcal{T}_\rho^{(m)}(\cdot)$ denotes the augmentation operator for the m-th view, which applies one or two randomly selected transformations from edge removal, node-feature masking, and BFS-based subgraph sampling. For each augmented view, the classifier produces a predictive distribution over the two classes, $\mathbf{p}_i^{(m)} = \mathrm{softmax}(f_\theta(G_i^{(m)}))$. The corresponding view pseudo-label is assigned according to the predicted class as $\widehat{y}_i^{(m)} = \arg\max \mathbf{p}_i^{(m)}$. The consensus prediction for G_i is then obtained by averaging the predictive distributions across all augmented views, *i.e.*,

$$\mathbf{p}_i^{\text{intra}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{p}_i^{(m)}, \qquad \widehat{y}_i^{\text{intra}} = \arg\max \mathbf{p}_i^{\text{intra}}, \tag{1}$$

where $\widehat{y}_i^{\text{intra}}$ is the corresponding consensus label. To characterize prediction consistency across augmented views, we further examine them at multiple subgroup granularities. Let $\mathcal{S}$ denote the set of candidate granularities, for a given granularity s, the M augmented views are randomly partitioned into $J_s = M/s$ equal-sized subgroups, and the partitioning is repeated to reduce the influence of random grouping. Let $\mathcal{J}_{i,j}^{(s)}$ denote the j-th subgroup under granularity s. The subgroup label $\widehat{y}_{i,j}^{(s)}$ is then obtained by selecting the class with the highest averaged prediction within each subgroup. For each granularity s, we evaluate the resulting subgroup predictions from two complementary aspects, *i.e.* within-subgroup coherence and cross-subgroup inconsistency. Specifically, we define

$$\eta_i(s) = \mathrm{Avg}\left(\mathbb{I}\left[\widehat{y}_i^{(m)} = \widehat{y}_{i,j}^{(s)}\right]\right), \qquad \phi_i(s) = \mathrm{Avg}\left(H\left(\mathbf{r}_i^{(s)}\right)\right), \tag{2}$$

At each granularity s, $\mathbf{r}_i^{(s)}$ denotes the distribution of subgroup labels aggregated over the repeated partitions, and $H(\cdot)$ denotes the normalized binary entropy. A larger $\eta_i(s)$ indicates stronger agreement between individual views and their subgroup decisions, whereas a smaller $\phi_i(s)$ indicates lower disagreement among subgroup predictions. Since improving within-subgroup coherence does not necessarily reduce cross-subgroup inconsistency, we retain the non-dominated candidates under Pareto dominance (Deb et al., 2002), forming $\mathcal{F}_i \subseteq \mathcal{S}$. The granularity closest to the ideal combination of high coherence and low inconsistency is then selected as

$$\hat{s}_i = \arg\min_{s \in \mathcal{F}_i} \sqrt{\lambda\big(1 - \eta_i(s)\big)^2 + (1 - \lambda)\phi_i(s)^2}, \tag{3}$$

where $\lambda \in [0, 1]$ balances the two criteria. The selected granularity $\hat{s}_i$ specifies the subgroup scale used to assess prediction consistency. We then measure the disagreement among subgroup predictions at this granularity without redefining the consensus prediction.

Definition 3.1. ***Perturbation Consistency Uncertainty.*** *For graph G_i, the perturbation consistency uncertainty at the selected granularity is defined as*

$$u_i^{\text{intra}} = H(\mathbf{p}_i^{\text{intra}}) - \mathrm{Avg}\left(H\left(\mathbf{p}_{i,j}^{(\hat{s}_i)}\right)\right), \tag{4}$$

where the averaging is taken over the subgroup predictions under the selected granularity $\hat{s}_i$ across repeated random partitions. A larger u_i^{intra} indicates greater distributional disagreement among the subgroup predictions around the resulting global consensus.

The following result formalizes this interpretation, with the proof provided in Appendix.

Theorem 3.1. *The perturbation consistency uncertainty in Eq. 4 is equivalent to the normalized generalized Jensen–Shannon divergence (Lin, 1991) among the subgroup predictions under the selected granularity $\hat{s}_i$:*

$$u_i^{\text{intra}} = \frac{1}{\log 2} \operatorname{Avg}\left(D_{\text{KL}}\left(\mathbf{p}_{i,j}^{(\hat{s}_i)} \,\middle\|\, \mathbf{p}_i^{\text{intra}}\right)\right), \tag{5}$$

where $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. Accordingly,

$$0 \le u_i^{\text{intra}} \le 1. \tag{6}$$

Furthermore, Pinsker's inequality (Kullback, 1967) yields

$$\operatorname{Avg}\left(\left\|\mathbf{p}_{i,j}^{(\hat{s}_i)} - \mathbf{p}_i^{\text{intra}}\right\|_1^2\right) \le 2\log 2\, u_i^{\text{intra}}. \tag{7}$$

Thus, a smaller u_i^{intra} implies tighter concentration of the subgroup predictions around the consensus distribution and hence stronger predictive consistency under graph perturbations.

3.3 Relational Propagation with Graph-of-Graphs Dependency Modeling

Although prediction consistency across augmented views provides evidence for assessing a graph, the prediction of an individual graph alone cannot guarantee the accuracy of its pseudo-label. Graph relationships provide complementary information by examining whether a prediction is supported by structurally similar graphs (Wang et al., 2022). Accordingly, we construct a graph-of-graphs and propagate class evidence among related graphs to validate the generated pseudo-labels.

For each graph G_i, let $\mathbf{z}_i^{\text{init}} \in \mathbb{R}^{d_z}$ denote its graph-level representation before optimization, where d_z denotes the dimensionality of the graph representation. We connect each graph to its k most similar training graphs via cosine similarity. This produces a binary adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$ and its row-normalized form $\widetilde{\mathbf{A}} = \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}_{i,i} = \sum_{j=1}^{N} A_{i,j}$. Evidence propagation requires class references. We use the known labelled samples as positive anchors, *i.e.*, $\mathcal{A}^+ = \mathcal{P}$, and select unlabeled graphs with the lowest positive-class probabilities predicted on the graphs as negative anchors $\mathcal{A}^-$. These anchors are used to initialize $\mathbf{E}^{(0)} \in \mathbb{R}^{N\times 2}$, where the corresponding class entry of each anchor is set to 1 and the remaining entries are set to 0. Starting from this initialization, class evidence is propagated over the graph-of-graphs by

$$\mathbf{E}^{(t+1)} = (1-\zeta)\widetilde{\mathbf{A}}\mathbf{E}^{(t)} + \zeta\mathbf{E}^{(0)}, \tag{8}$$

where t denotes the propagation iteration and $\zeta \in [0,1]$ controls anchor retention. Iteration continues until the propagated evidence becomes stable. The resulting evidence reflects the class supported by neighboring graphs. Following evidential learning, we represent the accumulated class evidence using a Dirichlet distribution (Sensoy et al., 2018), where one is added to each class-specific evidence value to obtain the Dirichlet parameters $\boldsymbol{\alpha}_i$. We then define

$$\mathbf{p}_i^{\text{inter}} = \frac{\boldsymbol{\alpha}_i}{\|\boldsymbol{\alpha}_i\|}, \qquad u_i^{\text{inter}} = \frac{2}{\|\boldsymbol{\alpha}_i\|}, \tag{9}$$

where $\mathbf{p}_i^{\text{inter}}$ gives the class distribution supported by the propagated evidence, and u_i^{inter} decreases as the evidence strength increases. We define the inter-graph prediction as $\widehat{y}_i^{\text{inter}} = \arg\max \mathbf{p}_i^{\text{inter}}$.

3.4 Complementary Evidence Integration for Reliable Pseudo-Labeling

After assessing prediction consistency across augmented views and validating predictions through relational propagation, we integrate the two forms of evidence to identify reliable pseudo-labels.

Since the two uncertainty values u_i^{intra} and u_i^{inter} have different scales, they are independently normalized using robust percentile-based scaling over the training graphs to obtain $\widetilde{u}_i^{\text{intra}}$ and $\widetilde{u}_i^{\text{inter}}$. Their corresponding reliabilities are $\xi_i^{\text{intra}} = 1 - \widetilde{u}_i^{\text{intra}}$ and $\xi_i^{\text{inter}} = 1 - \widetilde{u}_i^{\text{inter}}$. We then use these reliabilities to combine the two predictions, *i.e.*,

$$q_i = \frac{2\xi_i^{\text{intra}}\xi_i^{\text{inter}}}{\xi_i^{\text{intra}} + \xi_i^{\text{inter}} + \varepsilon}, \qquad \mathbf{p}_i^{\text{joint}} = \frac{\xi_i^{\text{intra}}\mathbf{p}_i^{\text{intra}} + \xi_i^{\text{inter}}\mathbf{p}_i^{\text{inter}}}{\xi_i^{\text{intra}} + \xi_i^{\text{inter}}}, \tag{10}$$

where q_i is the joint confidence and $\varepsilon > 0$ is a numerical constant. This joint confidence becomes small when either reliability is weak, whereas $\mathbf{p}_i^{\text{joint}}$ gives influence to the more reliable prediction. From the fused distribution, we obtain $\widehat{y}_i^{\text{joint}} = \arg\max \mathbf{p}_i^{\text{joint}}$ and the probability margin $\Delta_i = \mathbf{p}_{i,(1)}^{\text{joint}} - \mathbf{p}_{i,(2)}^{\text{joint}}$, where $\mathbf{p}_{i,(1)}^{\text{joint}}$ and $\mathbf{p}_{i,(2)}^{\text{joint}}$ denote the largest and second-largest class probabilities. Pseudo-label admission is restricted to graphs for which the two views agree and the fused decision is sufficiently reliable. For each class c, we form the class-specific candidate set $\mathcal{R}_c$,

$$\mathcal{R}_c = \left\{ i \in \mathcal{U} \;\middle|\; \widehat{y}_i^{\text{intra}} = \widehat{y}_i^{\text{inter}} = c,\ q_i \geq \tau,\ \Delta_i \geq \tau_\Delta \right\}, \tag{11}$$

where τ and τ_Δ denote the confidence and margin thresholds. The resulting high-confidence candidates in each $\mathcal{R}_c$ are retained as pseudo-labeled samples for subsequent training.

3.5 Geometric Deviance Estimation for Noise Elimination

The preceding complementary estimation provides reliable pseudo-labels for subsequent training, but some incorrect assignments may still remain and introduce noisy supervision during optimization. During training, mislabeled graphs may gradually deviate from the representation structure of their assigned classes (Yuan et al., 2026). Therefore, we further construct class prototypes to directly measure the geometric deviation of pseudo-labeled graphs from their assigned class structure and eliminate noisy pseudo-labels from subsequent supervision.

To begin with, we assess the initial geometric deviation of retained candidates from their assigned class structure using an anchor-based prototype constraint. For each class, an initial prototype is constructed from its anchor graphs as $\boldsymbol{\mu}_c^{\text{init}} = \sum_{a \in \mathcal{A}_c} \omega_a \mathbf{z}_a^{\text{init}} / \sum_{a \in \mathcal{A}_c} \omega_a$, where $\mathcal{A}_1 = \mathcal{A}^+$, $\mathcal{A}_0 = \mathcal{A}^-$, and $\omega_a = (1 - q_a + \varepsilon)^{-1}$. A retained candidate is admitted to class c only if The admitted candidate set for class c is defined as

$$\mathcal{U}_c = \left\{ i \in \mathcal{R}_c \;\middle|\; \left\|\mathbf{z}_i^{\text{init}} - \boldsymbol{\mu}_c^{\text{init}}\right\|_2 < \left\|\mathbf{z}_i^{\text{init}} - \boldsymbol{\mu}_{1-c}^{\text{init}}\right\|_2,\ \left\|\mathbf{z}_i^{\text{init}} - \boldsymbol{\mu}_c^{\text{init}}\right\|_2 \leq \tau_c^{\text{init}} \right\}, \tag{12}$$

where $\boldsymbol{\mu}_c^{\text{init}}$ denotes the prototype of class c in the initial representation space, and τ_c^{init} is the corresponding class-specific distance threshold. The prototype-distance comparison removes candidates that are closer to the opposite class, while the radius constraint excludes those lying too far from the prototype of their assigned class. Graphs that do not satisfy these requirements remain unlabeled, while labeled positive graphs retain their original labels. The resulting sets $\mathcal{U}_0$ and $\mathcal{U}_1$ are used as pseudo-label supervision in the subsequent optimization stage. At the late training stage, we further re-evaluate the geometric deviation of the retained pseudo-labeled graphs using their updated representations. Let $\mathbf{z}_i^{\text{late}}$ denote the learned representation of G_i at the late training stage. For each candidate $i \in \mathcal{U}_c$, the class reference set $\mathcal{Q}_{c,i}^{\text{late}}$ contains the original anchors and the other pseudo-labeled graphs assigned to class c. This leave-one-out construction prevents the representation of G_i from contributing to the class prototype used to evaluate its consistency. The class prototype and the deviation of G_i from that prototype are computed as

$$\boldsymbol{\mu}_{c,i}^{\text{late}} = \frac{\sum_{h \in \mathcal{Q}_{c,i}^{\text{late}}} \omega_h \mathbf{z}_h^{\text{late}}}{\sum_{h \in \mathcal{Q}_{c,i}^{\text{late}}} \omega_h}, \qquad d_i^{\text{late}} = \left\|\mathbf{z}_i^{\text{late}} - \boldsymbol{\mu}_{c,i}^{\text{late}}\right\|_2, \tag{13}$$

where $\boldsymbol{\mu}_{c,i}^{\text{late}}$ denotes the class prototype constructed from $\mathcal{Q}_{c,i}^{\text{late}}$, and d_i^{late} denotes the Euclidean distance between G_i and the corresponding prototype in the learned representation space. A larger d_i^{late} indicates greater inconsistency between the pseudo-labeled graph and its assigned class. Since the two classes may exhibit different within-class variation, a single threshold may not adequately reflect their class-specific deviation distributions. Therefore, We determine τ_c^{late} separately for each class as a predefined quantile of $d_i^{\text{late}} \mid i \in \mathcal{U}_c$. Pseudo-labeled graphs whose late-stage deviations exceed the corresponding thresholds are collected as

$$\mathcal{W} = \left\{ i \in \mathcal{U}_c \;\middle|\; d_i^{\text{late}} > \tau_c^{\text{late}} \right\}, \tag{14}$$

where $\mathcal{W}$ denotes the set of pseudo-labeled graphs whose deviations exceed the thresholds of their assigned classes. The final retained set is $\mathcal{U}_{\text{final}} = (\mathcal{U}_0 \cup \mathcal{U}_1) \setminus \mathcal{W}$, containing the pseudo-labeled graphs that satisfy both the initial reliability screening and the late-stage geometric criterion.

Table 1: Macro-F1 and Accuracy (%) under the observed-positive ratio $r_{\text{obs}} = 0.1$. Arrows indicate performance differences relative to NNRE, and the best results are highlighted in bold.

Model	MUTAG		NCI1		NCI109		PROTEINS		BAMultiShapes	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
NNRE	65.34	**73.98**	50.36	54.89	51.00	54.20	49.15	61.73	68.40	70.30
AttenPool	60.72 $_{\downarrow 04.62}$	72.31 $_{\downarrow 01.67}$	51.29 $_{\uparrow 00.93}$	54.38 $_{\downarrow 00.51}$	51.41 $_{\uparrow 00.41}$	54.66 $_{\uparrow 00.46}$	49.15 $_{\uparrow 0.00}$	61.37 $_{\downarrow 00.36}$	72.54 $_{\uparrow 04.14}$	73.80 $_{\uparrow 03.50}$
AddPool	59.20 $_{\downarrow 06.14}$	66.52 $_{\downarrow 07.46}$	49.59 $_{\downarrow 00.77}$	52.87 $_{\downarrow 02.02}$	52.27 $_{\uparrow 01.27}$	54.13 $_{\downarrow 00.07}$	54.14 $_{\uparrow 04.99}$	62.73 $_{\uparrow 01.00}$	68.03 $_{\downarrow 00.37}$	70.60 $_{\uparrow 00.30}$
MaxPool	63.26 $_{\downarrow 02.08}$	71.35 $_{\downarrow 02.63}$	53.61 $_{\uparrow 03.25}$	56.11 $_{\uparrow 01.22}$	53.75 $_{\uparrow 02.75}$	54.88 $_{\uparrow 00.68}$	46.31 $_{\downarrow 02.84}$	61.37 $_{\downarrow 00.36}$	65.00 $_{\downarrow 03.40}$	67.30 $_{\downarrow 03.00}$
GMTPool	60.96 $_{\downarrow 04.38}$	70.20 $_{\downarrow 03.78}$	33.31 $_{\downarrow 17.05}$	49.95 $_{\downarrow 04.94}$	33.17 $_{\downarrow 17.83}$	49.63 $_{\downarrow 04.57}$	40.86 $_{\downarrow 08.29}$	60.29 $_{\downarrow 01.44}$	61.40 $_{\downarrow 07.00}$	68.90 $_{\downarrow 01.40}$
URE	41.48 $_{\downarrow 23.86}$	67.02 $_{\downarrow 06.96}$	43.48 $_{\downarrow 06.88}$	51.21 $_{\downarrow 03.68}$	45.21 $_{\downarrow 05.79}$	51.66 $_{\downarrow 02.54}$	47.29 $_{\downarrow 01.86}$	57.68 $_{\downarrow 04.05}$	51.07 $_{\downarrow 17.33}$	61.20 $_{\downarrow 09.10}$
WCE	40.97 $_{\downarrow 24.37}$	45.09 $_{\downarrow 28.89}$	47.41 $_{\downarrow 02.95}$	52.87 $_{\downarrow 02.02}$	49.18 $_{\downarrow 01.82}$	54.23 $_{\uparrow 00.03}$	41.27 $_{\downarrow 07.88}$	60.38 $_{\downarrow 01.35}$	49.37 $_{\downarrow 19.03}$	57.30 $_{\downarrow 13.00}$
DELTA-PU	69.75 $_{\uparrow 04.41}$	71.40 $_{\downarrow 02.58}$	61.95 $_{\uparrow 11.59}$	**62.46** $_{\uparrow 07.57}$	62.43 $_{\uparrow 11.43}$	62.66 $_{\uparrow 08.46}$	57.34 $_{\uparrow 08.19}$	62.27 $_{\uparrow 00.54}$	83.08 $_{\uparrow 14.68}$	83.20 $_{\uparrow 12.90}$
Dist-PU	39.92 $_{\downarrow 25.42}$	66.49 $_{\downarrow 07.49}$	43.08 $_{\downarrow 07.28}$	52.89 $_{\downarrow 02.00}$	38.21 $_{\downarrow 12.79}$	51.73 $_{\downarrow 02.47}$	41.47 $_{\downarrow 07.68}$	59.39 $_{\downarrow 02.34}$	70.29 $_{\uparrow 01.89}$	72.30 $_{\uparrow 02.00}$
PEACE	**70.19** $_{\uparrow 04.85}$	72.26 $_{\downarrow 01.72}$	**62.16** $_{\uparrow 11.80}$	62.38 $_{\uparrow 07.49}$	**62.64** $_{\uparrow 11.64}$	**63.19** $_{\uparrow 08.99}$	**61.28** $_{\uparrow 12.13}$	**63.98** $_{\uparrow 02.25}$	**83.50** $_{\uparrow 15.10}$	**83.70** $_{\uparrow 13.40}$

3.6 SUMMARIZATION

In practice, **PEACE** optimizes the classifier using verified positives, retained pseudo-labels, and the remaining unlabeled graphs. For a training target $y \in \{0, 1\}$, the graph-level cross-entropy loss is $\ell(G_i, y) = -\log[\text{softmax}(f_\theta(G_i))]_y$. The overall training objective is defined as $\ell(G_i, y) = -\log[\text{softmax}(f_\theta(G_i))]_y$. The overall training objective is

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \ell(G_i, 1) + \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}_{\text{final}}} q_i \, \ell(G_i, \widehat{y}_i^{\text{joint}}) + \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U} \setminus (\mathcal{U}_{\text{final}} \cup \mathcal{W})} \ell(G_i, 0), \tag{15}$$

where q_i weights each retained pseudo-label according to its joint confidence, while graphs in $\mathcal{W}$ are removed from subsequent class supervision.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. To comprehensively evaluate **PEACE** under the positive–unlabeled (PU) graph classification setting, we conduct experiments on five benchmark datasets spanning molecular, bioinformatics, and synthetic domains. *MUTAG*, *NCI1*, *NCI109*, and *PROTEINS* are standard graph classification benchmarks from TUDataset (Morris et al., 2020), where the first three consist of molecular graphs and *PROTEINS* contains protein graphs. *BAMultiShapes* is a synthetic dataset composed of Barabási–Albert graphs with multiple structural motifs (Azzolin et al., 2022). Following the standard PU protocol, we use stratified 10-fold splits and retain only a fraction r_{obs} of positive training graphs as labeled positives, while treating all remaining training graphs as unlabeled. We consider $r_{\text{obs}} \in \{0.1, 0.2, 0.3\}$ to evaluate robustness under different supervision levels. Test labels are kept intact and never used for training or pseudo-label construction.

Baselines. We compare **PEACE** with representative baselines from two groups: graph representation methods and PU learning methods. The former include ADDPOOL (Duvenaud et al., 2015), MAXPOOL, ATTENPOOL (Meltzer et al., 2020), and GMTPOOL (Baek et al., 2021), covering commonly used graph-level aggregation strategies. The PU learning baselines include WCE, URE (Du Plessis et al., 2015), NNRE (Kiryo et al., 2017), and DIST-PU (Zhao et al., 2022), covering representative risk-estimation and distribution-based PU learning approaches. We additionally include DELTA-PU (Kim et al., 2026), a recent method specifically developed for PU graph classification. All methods are evaluated using identical data splits and observed-positive ratios under the same experimental protocol for fair comparison.

Implementation Details and Evaluation Metrics. We adopt a 5-layer GIN (Xu et al., 2018) with 64 hidden dimensions and mean pooling as the graph encoder. For **PEACE**, the key hyperparameters are set to $\rho = 0.1$, $k = 3$, and the pseudo-label confidence threshold $\tau = 0.8$ by default. We evaluate all methods using Macro-F1 and Accuracy (Acc), with detailed implementation settings and evaluation protocols provided in the Appendix.

Table 2: Macro-F1 and Accuracy (%) under the observed-positive ratio $r_{\text{obs}} = 0.2$. Arrows indicate performance differences relative to NNRE, and the best results are highlighted in bold.

Model	MUTAG		NCI1		NCI109		PROTEINS		BAMultiShapes	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
NNRE	66.34	75.53	49.25	54.89	47.68	52.03	49.18	60.83	70.05	71.60
AttenPool	60.09$_{\downarrow 06.25}$	72.34$_{\downarrow 03.19}$	48.70$_{\downarrow 00.55}$	54.35$_{\downarrow 00.54}$	51.44$_{\uparrow 03.76}$	55.32$_{\uparrow 03.29}$	50.55$_{\uparrow 01.37}$	61.82$_{\uparrow 00.99}$	67.96$_{\downarrow 02.09}$	70.40$_{\downarrow 01.20}$
AddPool	64.16$_{\downarrow 02.18}$	71.84$_{\downarrow 03.69}$	49.77$_{\uparrow 00.52}$	54.79$_{\downarrow 00.10}$	49.84$_{\uparrow 02.16}$	54.28$_{\uparrow 02.25}$	46.17$_{\downarrow 03.01}$	59.83$_{\downarrow 01.00}$	69.18$_{\downarrow 00.87}$	71.30$_{\downarrow 00.30}$
MaxPool	62.72$_{\downarrow 03.62}$	72.25$_{\downarrow 03.28}$	51.54$_{\uparrow 02.29}$	55.25$_{\uparrow 00.36}$	54.94$_{\uparrow 07.26}$	56.21$_{\uparrow 04.18}$	50.80$_{\uparrow 01.62}$	61.90$_{\uparrow 01.07}$	74.48$_{\uparrow 04.43}$	75.80$_{\uparrow 04.20}$
GMTPool	49.62$_{\downarrow 16.72}$	63.22$_{\downarrow 12.31}$	33.31$_{\downarrow 15.94}$	49.95$_{\downarrow 04.94}$	33.17$_{\downarrow 14.51}$	49.63$_{\downarrow 02.40}$	38.46$_{\downarrow 10.72}$	59.75$_{\downarrow 01.08}$	63.29$_{\downarrow 06.76}$	70.20$_{\downarrow 01.40}$
URE	43.34$_{\downarrow 23.00}$	68.07$_{\downarrow 07.46}$	47.46$_{\downarrow 01.79}$	53.23$_{\downarrow 01.66}$	46.71$_{\downarrow 00.97}$	52.36$_{\uparrow 00.33}$	44.11$_{\downarrow 05.07}$	59.75$_{\downarrow 01.08}$	53.77$_{\downarrow 16.28}$	62.70$_{\downarrow 08.90}$
WCE	48.00$_{\downarrow 18.34}$	49.88$_{\downarrow 25.65}$	48.66$_{\downarrow 00.59}$	54.04$_{\downarrow 00.85}$	45.15$_{\downarrow 02.53}$	52.66$_{\uparrow 00.63}$	38.52$_{\downarrow 10.66}$	59.39$_{\downarrow 01.44}$	64.34$_{\downarrow 05.71}$	67.20$_{\downarrow 04.40}$
DELTA-PU	67.87$_{\uparrow 01.53}$	71.32$_{\downarrow 04.21}$	62.91$_{\uparrow 13.66}$	63.60$_{\uparrow 08.71}$	62.93$_{\uparrow 15.25}$	63.70$_{\uparrow 11.67}$	61.35$_{\uparrow 12.17}$	64.15$_{\uparrow 03.32}$	**88.68**$_{\uparrow 18.63}$	**88.70**$_{\uparrow 17.10}$
Dist-PU	39.92$_{\downarrow 26.42}$	66.49$_{\downarrow 09.04}$	46.14$_{\downarrow 03.11}$	54.23$_{\downarrow 00.66}$	35.63$_{\downarrow 12.05}$	50.64$_{\downarrow 01.39}$	41.81$_{\downarrow 07.37}$	59.93$_{\downarrow 00.90}$	75.77$_{\uparrow 05.72}$	77.70$_{\uparrow 06.10}$
PEACE	**73.82**$_{\uparrow 07.48}$	**76.64**$_{\uparrow 01.11}$	**63.30**$_{\uparrow 14.05}$	**63.70**$_{\uparrow 08.81}$	**63.17**$_{\uparrow 15.49}$	**63.92**$_{\uparrow 11.89}$	**61.47**$_{\uparrow 12.29}$	**64.88**$_{\uparrow 04.05}$	88.58$_{\uparrow 18.53}$	88.60$_{\uparrow 17.00}$

4.2 EMPIRICAL RESULTS

Overall Performance. Tables 1 and 2 report the main comparison results on five graph classification datasets under $r_{\text{obs}} = 0.1$ and $r_{\text{obs}} = 0.2$, while Figure 2 presents results on NCI1 and PROTEINS under $r_{\text{obs}} = 0.3$. Overall, **PEACE** achieves the best performance in 16 of the 20 dataset–metric settings under $r_{\text{obs}} \in \{0.1, 0.2\}$, demonstrating consistent advantages across datasets and supervision levels. ❶ ***PEACE substantially outperforms NNRE across the evaluated datasets in both Macro-F1 and Accuracy.*** Specifically, **PEACE** improves Macro-F1 and Accuracy by 12.34% and 7.33% on average, with clear gains on NCI1, NCI109, PROTEINS, and BAMultiShapes. These results demonstrate that the proposed reliability estimation and noise elimination effectively improve classification performance from unlabeled graphs, especially under limited positive supervision. ❷ ***PEACE also maintains a consistent advantage over the stronger graph-specific baseline Delta-PU.*** It achieves average improvements of 1.18% in Macro-F1 and 0.98% in Accuracy, outperforming DELTA-PU in 17 of the 20 pairwise comparisons. The gains are evident on PROTEINS under $r_{\text{obs}} = 0.1$ and MUTAG under $r_{\text{obs}} = 0.2$, supporting the benefit of combining complementary reliability information with dynamic noise elimination for more reliable graph-level learning. ❸ ***PEACE remains effective across different observed-positive ratios.*** It maintains competitive or superior performance under $r_{\text{obs}} = 0.1$ and 0.2, while the results under $r_{\text{obs}} = 0.3$ show a similar overall trend, indicating robustness to varying levels of positive supervision. The results are presented in the appendix.

Ablation Study. We introduce several variants of **PEACE** to analyze the contribution of each component: (1) **PEACE** w/o All removes all three designs; (2) **PEACE** w/o PCS removes the perturbation-consistency component; (3) **PEACE** w/o RDP removes the relational-propagation component; and (4) **PEACE** w/o GDE disables the geometric-deviance-based filtering mechanism. As shown in Table 3, the ablation results demonstrate that all three designs contribute to the overall performance of **PEACE**. As shown in Table 3, the three components contribute to **PEACE** from complementary perspectives. Among them, PCS has the largest impact, as removing it decreases the average Macro-F1 by 1.96%, highlighting the importance of perturbation consistency for reliable pseudo-label construction. Removing RDP results in an average decrease of 0.89%, showing that relational information provides complementary evidence beyond individual graph predictions. GDE yields a smaller average improvement of 0.34%, with its downstream effect varying across datasets and observed-positive ratios. To further examine its intended role beyond the final classification metric, we provide a direct pseudo-label quality analysis in Appendix 6, where GDE reduces the retained pseudo-label error rate on four of the five datasets and preferentially removes erroneous pseudo-labels in most cases. When all three components are removed, the average Macro-F1 decreases by 1.61%. Overall, these results show that PCS and RDP primarily strengthen pseudo-label construction and calibration, while GDE provides an additional mechanism for suppressing unreliable pseudo-labels during subsequent learning.

Table 3: Macro-F1 (%) ablation study under different observed positive ratios $r_{\text{obs}} = 0.1/0.2$.

Method	NCI1	MUTAG	BAMultiShapes
r_{obs}	0.1/0.2	0.1/0.2	0.1/0.2
PEACE w/o All	59.76/62.02	67.13/74.33	80.79/87.87
PEACE w/o PCS	59.77/63.46	64.02/72.05	84.45/86.04
PEACE w/o RDP	59.90/61.23	70.58/73.24	84.33/86.96
PEACE w/o GDE	62.29/63.53	69.36/71.32	85.55/87.46
PEACE	**62.16/63.30**	**70.19/73.82**	**83.50/88.58**

Sensitivity Analysis. We further evaluate the sensitivity of **PEACE** to the scope weight λ and the joint threshold τ. As shown in Figure 2(a), varying λ from 0.3 to 0.9 causes only minor performance fluctuations across most datasets, indicating that **PEACE** is relatively insensitive to the weighting

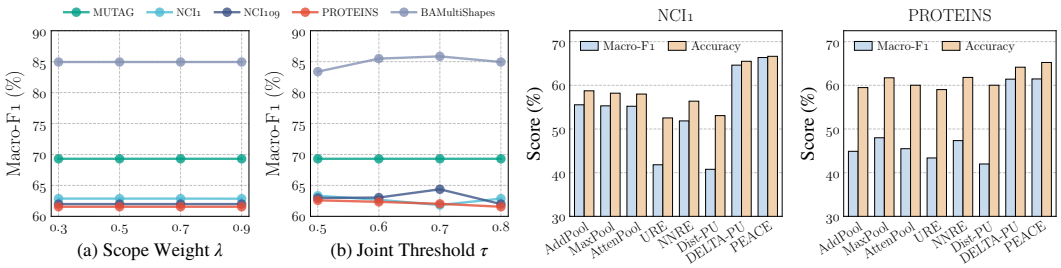


Figure 2: Sensitivity analysis of **PEACE** under different λ and τ, and model comparison on NCI1 and PROTEINS. (a)(b) show the Macro-F1 trends under varying parameters. The right panels compare Macro-F1 and Accuracy across baselines, where **PEACE** achieves the best overall performance.

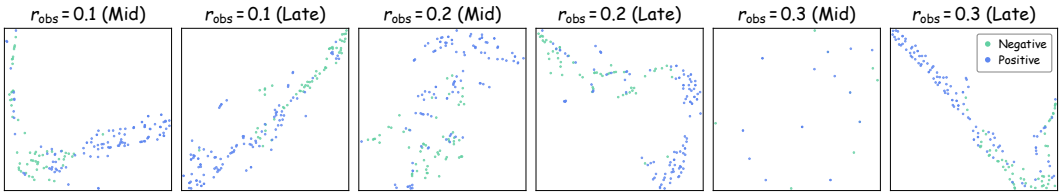


Figure 3: t-SNE visualization of graph embeddings on MUTAG under $r_{\text{obs}} \in \{0.1, 0.2, 0.3\}$ at the mid and late training stages.Green and blue dots represent negative and positive samples, respectively, illustrating the evolution of feature representations and class separability.

of different evidence. Moderate values generally maintain stable performance, whereas strongly biased weighting brings no consistent improvement. Figure 2(b) shows that the influence of τ is more dataset-dependent. MUTAG remains nearly unchanged across different threshold values, while NCI109 and BAMultiShapes generally achieve better performance with moderate thresholds. In contrast, excessively large τ values can degrade performance on NCI1 and PROTEINS because overly strict selection may exclude informative unlabeled graphs. Overall, both λ and τ exhibit relatively broad stable ranges across datasets, showing that **PEACE** does not require fine-grained dataset-specific hyperparameter tuning.

Embedding Visualization. Figure 3 visualizes t-SNE projections of **PEACE** embeddings on MUTAG for $r_{\text{obs}} \in \{0.1, 0.2, 0.3\}$. Compared with the mid-stage representations, the late-stage embeddings generally exhibit more compact and class-aware structures across different observed-positive ratios, with reduced intra-class dispersion and clearer class organization. This trend indicates that **PEACE** progressively suppresses task-irrelevant noise and learns more discriminative graph representations during training. The improvement is particularly evident at $r_{\text{obs}} = 0.1$, where the two classes form more coherent regions with reduced overlap despite limited positive supervision. For $r_{\text{obs}} = 0.2$ and 0.3, the embeddings also become more structured, although noticeable overlap remains between structurally similar graphs. Overall, these results show that **PEACE** progressively learns more compact and discriminative graph representations across different supervision levels.

5 Conclusion

This paper presents **PEACE**, a framework for positive-unlabeled graph classification. By combining perturbation consistency with relational propagation, **PEACE** identifies reliable supervision signals from unlabeled graphs and integrates evidence to improve pseudo-label quality under the absence of negative graphs. To further mitigate the influence of noisy supervision, **PEACE** exploits the geometric structure of the latent space to dynamically identify and eliminate unreliable graph samples according to their prototypical deviance, thereby reducing error accumulation during optimization. Extensive experiments on multiple graph classification benchmarks demonstrate that **PEACE** achieves competitive performance over state-of-the-art PU graph classification approaches, validating the effectiveness of the proposed evidence certification and geometric elimination mechanisms.

AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation and code development. We did not use generative AI tools for research ideation, hypothesis formulation, methodology or experiment design, theoretical development, or the analysis and interpretation of experimental results. Additionally, generative AI tools were used for language editing, including grammar correction, limited rephrasing, and improvements to readability. All AI-assisted code was manually reviewed and tested for correctness, and all AI-assisted language edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only publicly available benchmark datasets and does not involve new data collection or direct interaction with human subjects. We are not aware of any specific ethical concerns arising from the methodology or experiments presented in this work. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the download links for all publicly available datasets used in our experiments, together with detailed data preprocessing instructions. The model architecture, training settings, hyperparameters, and experimental protocols are described in the main paper and appendix. The complete implementation and model code will be publicly released upon publication of this work. An anonymized code repository is available at https://anonymous.4open.science/r/Anonymous_code-36EA.

REFERENCES

Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. Global explainability of gnns via logic combination of learned concepts. 2022.

Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. 2021.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2): 182–197, 2002.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pp. 1386–1394. PMLR, 2015.

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.

Changda Gong, Jiaojiao Fang, Guixia Liu, Yun Tang, and Weihua Li. Ulcyp: A multitask model for predicting p450 inducers based on positive-unlabeled learning. *Journal of Chemical Information and Modeling*, 66(10):5838–5850, 2026.

Wei Ju, Xiao Luo, Meng Qu, Yifan Wang, Chong Chen, Minghua Deng, Xian-Sheng Hua, and Ming Zhang. Tgnn: A joint semi-supervised framework for graph-level classification. 2023.

Wei Ju, Zhengyang Mao, Ziyue Qiao, Yifang Qin, Siyu Yi, Zhiping Xiao, Xiao Luo, Yanjie Fu, and Ming Zhang. Focus on informative graphs! semi-supervised active learning for graph-level classification. *Pattern Recognition*, 153:110567, 2024a.

Wei Ju, Zhengyang Mao, Siyu Yi, Yifang Qin, Yiyang Gu, Zhiping Xiao, Yifan Wang, Xiao Luo, and Ming Zhang. Hypergraph-enhanced dual semi-supervised graph classification. 2024b.

Junghun Kim, Shihyung Park, and U Kang. Dual-level reweighting for positive-unlabeled graph classification. In *Proceedings of the ACM Web Conference 2026*, pp. 834–844, 2026.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. volume 30, 2017.

Eunji Koh, Young Jae Lee, and Seoung Bum Kim. Graphixmatch: Improving semi-supervised learning for graph classification with fixmatch. *Neurocomputing*, 607:128356, 2024.

Solomon Kullback. A lower bound for discrimination information in terms of variation (corresp.). *IEEE transactions on Information Theory*, 13(1):126–127, 1967.

Atsutoshi Kumagai, Tomoharu Iwata, Hiroshi Takahashi, Taishi Nishiyama, and Yasuhiro Fujiwara. Importance-weighted positive-unlabeled learning for distribution shift adaptation. In *AISTATS*, pp. 1576–1584, 2025.

Chunquan Liang, Yi Tian, Dongmin Zhao, Mei Li, Shirui Pan, Hongming Zhang, and Jicheng Wei. Bootstrap latent prototypes for graph positive-unlabeled learning. *Information Fusion*, 112: 102553, 2024.

Chunquan Liang, Luyue Wang, Xinyuan Feng, Yuying Cheng, Mei Li, Shirui Pan, and Hongming Zhang. Graph positive-unlabeled learning via bootstrapping label disambiguation. *Neural Networks*, 190:107630, 2025.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Bo Liu, Tiantian Peng, Yanshan Xiao, Shilei Zhao, Peng Sun, Xiaokai Li, Zhiyu Zheng, and Yongsheng Huang. Self-paced multi-view positive and unlabeled graph learning with auxiliary information. *Information Sciences*, 642:119146, 2023.

Bo Liu, Chenlong Ye, Yanshan Xiao, Baoqing Li, Zhitong Wang, Boxu Zhou, Shengxin He, and Fan Cao. A multi-view forward positive and unlabeled graph learning method based on dictionary learning. *Information Sciences*, pp. 122517, 2025.

Zhengyang Liu, Hang Yu, and Xiangfeng Luo. Mitigating label noise in graph learning with information bottleneck-guided aggregation and adversarial consistency. *Pattern Recognition*, pp. 114271, 2026.

Peter Meltzer, Marcelo Daniel Gutierrez Mallea, and Peter J Bentley. Pinet: Attention pooling for graph classification. *arXiv preprint arXiv:2008.04575*, 2020.

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.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. volume 31, 2018.

Alex Smola. Protein function prediction via graph kernels. *Bioinformatics*, 2005.

Xusheng Sun, Jingli Huang, Gang Wang, Xuan Zhang, Manlu Liu, and Shanlin Yang. Uncovering financial statement fraud detection with multi-modal positive-unlabeled contrastive learning. *Electronic Commerce Research and Applications*, pp. 101604, 2026.

Nikil Wale, Ian A Watson, and George Karypis. Comparison of descriptor spaces for chemical compound retrieval and classification. *Knowledge and Information Systems*, 14(3):347–375, 2008.

Yifan Wang, Xiao Luo, Chong Chen, Xian-Sheng Hua, Ming Zhang, and Wei Ju. Disensemi: Semi-supervised graph classification via disentangled representation learning. *IEEE transactions on neural networks and learning systems*, 36(5):8192–8204, 2024.

Yu Wang, Yuying Zhao, Neil Shah, and Tyler Derr. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 2067–2076, 2022.

Jia Wu, Shirui Pan, Xingquan Zhu, Chengqi Zhang, and Xindong Wu. Positive and unlabeled multi-graph learning. *Ieee transactions on cybernetics*, 47(4):818–829, 2016.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? 2018.

Suqin Yuan, Lei Feng, Bo Han, and Tongliang Liu. Enhancing sample selection against label noise by cutting mislabeled easy examples. volume 38, pp. 43942–43970, 2026.

Yuchen Zhao, Xiangnan Kong, and S Yu Philip. Positive and unlabeled learning for graph classification. In *2011 IEEE 11th International Conference on Data Mining*, pp. 962–971. IEEE, 2011.

Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, and Qingming Huang. Dist-pu: Positive-unlabeled learning from a label distribution perspective. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14441–14450. IEEE, 2022.

Fernando Zhapa-Camacho, Zhenwei Tang, Maxat Kulmanov, and Robert Hoehndorf. Predicting protein functions using positive-unlabeled ranking with ontology-based priors. *Bioinformatics*, 40(Supplement_1):i401–i409, 2024.

A ALGORITHM

The complete training procedure of **PEACE** is summarized in Algorithm 1. The procedure consists of four successive stages corresponding to the main components of **PEACE**: perturbation consistency with subgroup granularity selection, relational propagation with graph-of-graphs dependency modeling, complementary evidence integration for reliable pseudo-labeling, and geometric deviance estimation for noise elimination.

Algorithm 1 Training Procedure of the Proposed **PEACE** Framework

Require: Training set $\mathcal{D} = \mathcal{P} \cup \mathcal{U}$, graph classifier f_θ, M, ρ, $\mathcal{S}$, λ, k, ζ, τ, and τ_Δ
Ensure: Trained model parameters θ

Obtain initial graph representations $\{\mathbf{z}_i^{\text{init}}\}_{i=1}^N$ and model predictions before pseudo-label construction.
Part I: Perturbation Consistency with Subgroup Granularity Selection
for each graph $G_i \in \mathcal{D}$ **do**
  Generate M augmented views $G_i^{(m)} = \mathcal{T}_\rho^{(m)}(G_i)$.
  Obtain predictive distributions $\mathbf{p}_i^{(m)}$ and view labels $\widehat{y}_i^{(m)}$.
  Compute the consensus prediction $\mathbf{p}_i^{\text{intra}}$ and label $\widehat{y}_i^{\text{intra}}$ (Eq. 1).
  for each granularity $s \in \mathcal{S}$ **do**
    Randomly partition the M views into $J_s = M/s$ subgroups with repeated partitions.
    Obtain subgroup predictions $\mathbf{p}_{i,j}^{(s)}$ and subgroup labels $\widehat{y}_{i,j}^{(s)}$.
    Compute coherence $\eta_i(s)$ and cross-subgroup inconsistency $\phi_i(s)$ (Eq. 2).
  end for
  Construct the Pareto set $\mathcal{F}_i$ and select the granularity $\hat{s}_i$ (Eq. 3).
  Compute the perturbation consistency uncertainty u_i^{intra} (Eq. 4).
end for
Part II: Relational Propagation with Graph-of-Graphs Dependency Modeling
Construct the kNN graph-of-graphs from $\{\mathbf{z}_i^{\text{init}}\}_{i=1}^N$.
Set $\mathcal{A}^+ = \mathcal{P}$, select negative anchors $\mathcal{A}^-$, and initialize the evidence matrix $\mathbf{E}^{(0)}$.
Propagate class evidence over the graph-of-graphs until convergence (Eq. 8).
Construct the Dirichlet parameters and obtain $\mathbf{p}_i^{\text{inter}}$, u_i^{inter}, and $\widehat{y}_i^{\text{inter}}$ (Eq. 9).
Part III: Complementary Evidence Integration for Reliable Pseudo-labeling
Normalize u_i^{intra} and u_i^{inter} and obtain the corresponding evidence reliabilities.
Fuse the intra- and inter-graph evidence to obtain q_i, $\mathbf{p}_i^{\text{joint}}$, $\widehat{y}_i^{\text{joint}}$, and Δ_i (Eq. 10).
Construct the class-specific candidate sets $\mathcal{R}_c$ according to prediction agreement and confidence criteria (Eq. 11).
Rank candidates in each $\mathcal{R}_c$ by q_i and retain candidates according to the class-wise selection rule.
Part IV: Geometric Deviance Estimation for Noise Elimination
Construct the initial class prototypes $\boldsymbol{\mu}_c^{\text{init}}$ from the corresponding anchor sets.
Apply the initial prototype-distance and radius constraints to obtain $\mathcal{U}_0$ and $\mathcal{U}_1$ (Eq. 12).
Optimize f_θ using the verified positives and admitted pseudo-labels to the late training stage.
Extract the late-stage graph representations $\{\mathbf{z}_i^{\text{late}}\}$.
for each pseudo-labeled graph $i \in \mathcal{U}_c$ **do**
  Construct $\mathcal{Q}_{c,i}^{\text{late}}$ and compute the leave-one-out prototype $\boldsymbol{\mu}_{c,i}^{\text{late}}$.
  Compute the geometric deviation d_i^{late} (Eq. 13).
end for
Determine the class-specific thresholds τ_c^{late} from the corresponding deviation distributions.
Construct the withdrawal set $\mathcal{W}$ (Eq. 14).
Obtain the final pseudo-labeled set $\mathcal{U}_{\text{final}} = (\mathcal{U}_0 \cup \mathcal{U}_1) \setminus \mathcal{W}$.
Continue optimizing f_θ with the refined supervision according to Eq. 15.
return final model parameters θ.

B PROOF OF THEOREM

This section provides the detailed proof of the theoretical result in the main text. Throughout the proof, log denotes the natural logarithm, and the normalized binary entropy is defined as

$$H(\mathbf{p}) = -\frac{1}{\log 2}\sum_{c=0}^{1} p_c \log p_c, \tag{16}$$

where $0 \log 0 = 0$.

Proof of Theorem 3.1. For graph G_i, consider the subgroup predictive distributions $\mathbf{p}_{i,j}^{(\hat{s}_i)}$ obtained under the selected granularity $\hat{s}_i$ across the repeated random partitions. Since each partition divides the M augmented views into equal-sized subgroups, averaging the subgroup distributions recovers the global consensus distribution in Eq. 1, namely,

$$\mathrm{Avg}\left(\mathbf{p}_{i,j}^{(\hat{s}_i)}\right) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{p}_i^{(m)} = \mathbf{p}_i^{\mathrm{intra}}. \tag{17}$$

By Definition 3.1,

$$u_i^{\mathrm{intra}} = H(\mathbf{p}_i^{\mathrm{intra}}) - \mathrm{Avg}\left(H\left(\mathbf{p}_{i,j}^{(\hat{s}_i)}\right)\right). \tag{18}$$

Expanding the normalized entropy gives

$$\begin{aligned} u_i^{\mathrm{intra}} = &-\frac{1}{\log 2}\sum_{c=0}^{1} p_{i,c}^{\mathrm{intra}} \log p_{i,c}^{\mathrm{intra}} \\ &+ \frac{1}{\log 2}\,\mathrm{Avg}\left(\sum_{c=0}^{1} p_{i,j,c}^{(\hat{s}_i)} \log p_{i,j,c}^{(\hat{s}_i)}\right). \end{aligned} \tag{19}$$

Using Eq. 17, the first term can be included inside the averaging operator, yielding

$$\begin{aligned} u_i^{\mathrm{intra}} &= \frac{1}{\log 2}\,\mathrm{Avg}\left(\sum_{c=0}^{1} p_{i,j,c}^{(\hat{s}_i)} \log \frac{p_{i,j,c}^{(\hat{s}_i)}}{p_{i,c}^{\mathrm{intra}}}\right) \\ &= \frac{1}{\log 2}\,\mathrm{Avg}\left(D_{\mathrm{KL}}\left(\mathbf{p}_{i,j}^{(\hat{s}_i)} \,\middle\|\, \mathbf{p}_i^{\mathrm{intra}}\right)\right). \end{aligned} \tag{20}$$

Therefore, u_i^{intra} is the normalized generalized Jensen–Shannon divergence among the subgroup predictions under the selected granularity $\hat{s}_i$, which establishes Eq. 5.

Since the Kullback–Leibler divergence is nonnegative,

$$u_i^{\mathrm{intra}} \geq 0. \tag{21}$$

Moreover, the normalized binary entropy satisfies $0 \leq H(\mathbf{p}) \leq 1$. Hence,

$$\begin{aligned} u_i^{\mathrm{intra}} &= H(\mathbf{p}_i^{\mathrm{intra}}) - \mathrm{Avg}\left(H\left(\mathbf{p}_{i,j}^{(\hat{s}_i)}\right)\right) \\ &\leq H(\mathbf{p}_i^{\mathrm{intra}}) \leq 1. \end{aligned} \tag{22}$$

Thus,

$$0 \leq u_i^{\mathrm{intra}} \leq 1. \tag{23}$$

Finally, Pinsker's inequality gives, for each subgroup prediction,

$$D_{\mathrm{KL}}\left(\mathbf{p}_{i,j}^{(\hat{s}_i)} \,\middle\|\, \mathbf{p}_i^{\mathrm{intra}}\right) \geq \frac{1}{2}\left\|\mathbf{p}_{i,j}^{(\hat{s}_i)} - \mathbf{p}_i^{\mathrm{intra}}\right\|_1^2. \tag{24}$$

Taking the average over the subgroup predictions yields

$$\begin{aligned} \mathrm{Avg}\left(\left\|\mathbf{p}_{i,j}^{(\hat{s}_i)} - \mathbf{p}_i^{\mathrm{intra}}\right\|_1^2\right) &\leq 2\,\mathrm{Avg}\left(D_{\mathrm{KL}}\left(\mathbf{p}_{i,j}^{(\hat{s}_i)} \,\middle\|\, \mathbf{p}_i^{\mathrm{intra}}\right)\right) \\ &= 2\log 2\, u_i^{\mathrm{intra}}, \end{aligned} \tag{25}$$

which proves Eq. 7.

Therefore, a smaller u_i^{intra} implies that the subgroup predictive distributions are more tightly concentrated around the global consensus distribution $\mathbf{p}_i^{\mathrm{intra}}$, completing the proof. □

C DETAILED EXPERIMENTAL SETTING

C.1 DATASET DETAILS

We conduct experiments on five benchmark graph classification datasets covering molecular, protein, and synthetic graphs.

MUTAG (Debnath et al., 1991). MUTAG is a molecular graph dataset consisting of 188 nitroaromatic compounds. Each graph represents a chemical compound, where nodes correspond to atoms and edges correspond to chemical bonds. Node features are 7-dimensional one-hot encoded vectors indicating atom types. The task is binary graph classification to predict whether a compound has mutagenic effect on *Salmonella typhimurium*.

NCI1 and NCI109 (Wale et al., 2008). NCI1 and NCI109 are two subsets of the National Cancer Institute (NCI) chemical compound dataset, containing 4,110 and 4,127 graphs, respectively. Each graph represents a chemical compound, with nodes as atoms and edges as chemical bonds. Node features are 37-dimensional vectors encoding atom types and other chemical attributes. The binary classification task is to predict whether a compound exhibits activity against non-small cell lung cancer (NCI1) or ovarian cancer (NCI109) cells.

PROTEINS (Smola, 2005). PROTEINS is a bioinformatics dataset of 1,113 protein graphs. Nodes represent secondary structure elements (helices, sheets, and turns), and edges connect elements that are spatially close or sequentially adjacent. Node features are 3-dimensional vectors encoding structural properties. The binary classification task is to predict whether a protein belongs to the enzyme or non-enzyme class.

BAMultiShapes (Azzolin et al., 2022). BAMultiShapes is a synthetic graph classification dataset designed for evaluating explainability algorithms. It contains 1,000 graphs, each generated by attaching structural motifs (house, wheel, and grid) to a random Barabási–Albert (BA) base graph. The dataset is pre-computed from the official implementation and the classification task is based on logical combinations of motif presence. Following the standard positive-unlabeled protocol in our experiments, the multi-class problem is binarized accordingly.

C.2 BASELINE DETAILS

In our experiments, we compare **PEACE** with nine representative baselines, which are categorized into two groups: graph representation learning methods and positive-unlabeled (PU) learning methods. For a fair comparison, all methods use the same data splits and observed-positive ratios, and are evaluated under the same graph-level PU setting. The details of each baseline are as follows.

ADDPOOL (Duvenaud et al., 2015). ADDPOOL constructs graph-level representations by summing node embeddings over the entire graph, providing a simple permutation-invariant aggregation strategy.

MAXPOOL. MAXPOOL performs dimension-wise maximum aggregation over node embeddings, retaining the most prominent features to form graph-level representations.

ATTENPOOL (Meltzer et al., 2020). ATTENPOOL assigns learnable attention weights to node representations and aggregates them according to their relative importance for graph-level prediction.

GMTPOOL (Baek et al., 2021). GMTPOOL employs a graph multiset transformer to model interactions among node representations and obtain expressive graph-level embeddings through attention-based pooling.

WCE. WCE applies class-dependent weights to the standard cross-entropy loss to alleviate the imbalance between labeled positive and unlabeled graphs.

URE (Du Plessis et al., 2015). URE reformulates the classification risk using positive and unlabeled data and derives an unbiased risk estimator without requiring labeled negative samples.

NNRE (Kiryo et al., 2017). NNRE introduces a non-negative correction to PU risk estimation, preventing the estimated negative risk from becoming negative and reducing overfitting.

Table 4: Macro-F1 and Accuracy (%) under the observed-positive ratio $r_{\text{obs}} = 0.3$. Arrows indicate performance differences relative to NNRE, and the best results are highlighted in bold.

Model	MUTAG		NCI1		NCI109		PROTEINS		BAMultiShapes	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
NNRE	65.41	74.94	51.85	56.37	46.39	54.23	47.34	61.81	75.04	75.80
AttenPool	65.36$_{\downarrow 0.05}$	75.00$_{\uparrow 0.06}$	55.21$_{\uparrow 03.36}$	58.00$_{\uparrow 01.63}$	49.97$_{\uparrow 03.58}$	55.08$_{\uparrow 0.85}$	45.49$_{\downarrow 01.85}$	60.02$_{\downarrow 01.79}$	67.20$_{\downarrow 07.84}$	69.80$_{\downarrow 06.00}$
AddPool	58.22$_{\downarrow 07.19}$	66.87$_{\downarrow 08.07}$	55.54$_{\uparrow 03.69}$	58.74$_{\uparrow 02.37}$	50.88$_{\uparrow 04.49}$	54.57$_{\uparrow 0.34}$	44.89$_{\downarrow 02.45}$	59.48$_{\downarrow 02.33}$	63.29$_{\downarrow 11.75}$	67.90$_{\downarrow 07.90}$
MaxPool	64.03$_{\downarrow 01.38}$	72.34$_{\downarrow 02.60}$	55.30$_{\uparrow 03.45}$	58.20$_{\uparrow 01.83}$	54.05$_{\uparrow 07.66}$	56.36$_{\uparrow 02.13}$	47.99$_{\uparrow 0.65}$	61.72$_{\downarrow 0.09}$	75.96$_{\uparrow 0.92}$	77.00$_{\uparrow 01.20}$
GMTPool	58.83$_{\downarrow 06.58}$	70.18$_{\downarrow 04.76}$	33.31$_{\downarrow 18.54}$	49.95$_{\downarrow 06.42}$	33.17$_{\downarrow 13.22}$	49.63$_{\downarrow 04.60}$	38.47$_{\downarrow 08.87}$	59.93$_{\downarrow 01.88}$	71.21$_{\downarrow 03.83}$	75.90$_{\uparrow 0.10}$
URE	42.41$_{\downarrow 23.00}$	67.54$_{\downarrow 07.40}$	41.81$_{\downarrow 10.04}$	52.53$_{\downarrow 03.84}$	42.27$_{\downarrow 04.12}$	51.06$_{\downarrow 03.17}$	43.37$_{\downarrow 03.97}$	59.02$_{\downarrow 02.79}$	51.84$_{\downarrow 23.20}$	61.60$_{\downarrow 14.20}$
WCE	56.55$_{\downarrow 08.86}$	56.96$_{\downarrow 17.98}$	51.73$_{\downarrow 0.12}$	55.74$_{\downarrow 0.63}$	48.59$_{\uparrow 02.20}$	54.74$_{\uparrow 0.51}$	40.04$_{\downarrow 07.30}$	59.66$_{\downarrow 02.15}$	64.20$_{\downarrow 10.84}$	67.70$_{\downarrow 08.10}$
DELTA-PU	71.44$_{\uparrow 06.03}$	77.11$_{\uparrow 02.17}$	64.62$_{\uparrow 12.77}$	65.50$_{\uparrow 09.13}$	65.91$_{\uparrow 19.52}$	66.47$_{\uparrow 12.24}$	61.41$_{\uparrow 14.07}$	64.15$_{\uparrow 02.34}$	**90.89**$_{\uparrow 15.85}$	**90.90**$_{\uparrow 15.10}$
Dist-PU	39.92$_{\downarrow 25.49}$	66.49$_{\downarrow 08.45}$	40.77$_{\downarrow 11.08}$	53.06$_{\downarrow 03.31}$	34.63$_{\downarrow 11.76}$	50.55$_{\downarrow 03.68}$	41.99$_{\downarrow 05.35}$	60.02$_{\downarrow 01.79}$	69.51$_{\downarrow 05.53}$	72.70$_{\downarrow 03.10}$
PEACE	**72.47**$_{\uparrow 07.06}$	**77.75**$_{\uparrow 02.81}$	**66.36**$_{\uparrow 14.51}$	**66.64**$_{\uparrow 10.27}$	**66.40**$_{\uparrow 20.01}$	**66.92**$_{\uparrow 12.69}$	**61.46**$_{\uparrow 14.12}$	**65.23**$_{\uparrow 03.42}$	90.58$_{\uparrow 15.54}$	90.60$_{\uparrow 14.80}$

DIST-PU (Zhao et al., 2022). DIST-PU approaches PU learning from a label-distribution perspective by encouraging consistency between predicted and expected label distributions for unlabeled samples.

DELTA-PU (Kim et al., 2026). DELTA-PU is designed for PU graph classification and employs dual-level reweighting to emphasize informative graph information while reducing the influence of ambiguous unlabeled graphs.

C.3 Implementation Settings

We use GIN as the default backbone for all experiments, with five graph layers, 64 hidden units, mean pooling, and a dropout rate of 0.5. The model is optimized using Adam with a learning rate of 0.01 and a batch size of 512 for 100 epochs at each training stage. The learning rate is decayed by a factor of 0.5 every 50 epochs. For graph relation validation, we set the subgroup trade-off coefficient to 0.7 and use $k = 3$ for graph-level neighborhood construction. For learning-guided label refinement, the joint confidence threshold and margin threshold are set to 0.8 and 0.2.

D More Experiments

D.1 Performance Extension Experiment

We further report the detailed results under the observed-positive ratio $r_{\text{obs}} = 0.3$. As shown in Table 4, **PEACE** achieves the best performance in 8 of the 10 dataset–metric settings, demonstrating that its advantage is maintained when more positive graphs are observed. Compared with the strongest baseline, DELTA-PU, **PEACE** improves both Macro-F1 and Accuracy by approximately 0.60% on average. The gains are particularly evident on NCI1, where **PEACE** improves Macro-F1 and Accuracy by 1.74% and 1.14%, respectively, while consistent improvements are also observed on MUTAG and NCI109. On PROTEINS, **PEACE** achieves comparable Macro-F1 while obtaining a clearer improvement in Accuracy. Although DELTA-PU slightly outperforms **PEACE** on BAMultiShapes, the performance gap remains within 0.31%. Overall, these results further confirm the effectiveness and robustness of **PEACE** under different observed-positive ratios.

D.2 Ablation Study.

Table 5 evaluates the five ablation variants at $r_{\text{obs}} = 0.2$. Averaged over the five datasets, the full model attains 70.07% Macro-F1. Compared with removing the intra module, the inter module, and all modules, the full model improves the average Macro-F1 by 1.37%, 1.15%, and 0.38%, respectively. The intra and inter removals therefore lead to the largest average performance degradation in this setting. Compared with the feedback removal, the full model improves the average Macro-F1 by 0.21%; however, the feedback removal achieves higher results on NCI1, NCI109, and PROTEINS. These results support the overall contribution of the intra/inter components, while the effect of the feedback component varies across datasets.

Table 5: Macro-F1 (%) ablation study on all datasets under different observed positive ratios $r_{\text{obs}} = 0.1/0.2/0.3$. Results in each cell are reported in the order $0.1/0.2/0.3$.

Method	**NCI1**	**NCI109**	**MUTAG**	**PROTEINS**	**BAMultiShapes**
r_{obs}	$0.1/0.2/0.3$	$0.1/0.2/0.3$	$0.1/0.2/0.3$	$0.1/0.2/0.3$	$0.1/0.2/0.3$
PEACE w/o All	59.76/62.02/66.01	61.31/62.34/65.22	67.13/74.33/67.79	61.55/61.85/61.71	80.79/87.87/88.47
PEACE w/o PCS	59.77/63.46/66.02	61.10/60.55/64.49	64.02/72.05/69.97	59.16/61.40/60.53	84.45/86.04/88.04
PEACE w/o RDP	59.90/61.23/63.05	58.34/62.75/63.36	70.58/73.24/72.78	63.05/60.38/57.90	84.33/86.96/88.37
PEACE w/o GDE	62.29/63.53/66.32	62.73/64.38/63.53	69.36/71.32/72.04	61.08/62.62/62.35	85.55/87.46/90.89
PEACE	**62.16/63.30/66.36**	**62.64/63.17/66.40**	**70.19/73.82/72.47**	**61.28/61.47/61.46**	**83.50/88.58/90.58**

Table 6: Diagnostic analysis of prototype-guided noise elimination at $r_{\text{obs}} = 0.2$. Error rates measure the proportion of incorrect pseudo-labels before and after elimination. "Wrong Removed" and "Correct Removed" denote the removal rates of incorrectly and correctly pseudo-labeled samples.

Dataset	Error Before ↓	Error Retained ↓	Wrong Removed ↑	Correct Removed ↓	Retention ↑
MUTAG	14.0	**4.3**	**92.9**	74.4	23.0
NCI1	34.3	**32.9**	**19.6**	14.4	83.8
NCI109	34.7	34.8	67.3	67.4	32.6
PROTEINS	34.7	**28.8**	**50.5**	34.8	59.7
BAMultiShapes	19.4	**15.5**	**34.4**	13.9	82.1
Average	27.4	**23.3**	**52.9**	41.0	56.2

D.3 Noise Elimination Analysis.

To further examine whether the proposed prototype-guided noise elimination effectively identifies unreliable pseudo-labels, we directly evaluate the quality of pseudo-labels before and after elimination. Specifically, we report the pseudo-label error rate before elimination (*Error Before*), the error rate among the retained samples (*Error Retained*), the removal rates of incorrectly and correctly pseudo-labeled samples (*Wrong Removed* and *Correct Removed*), and the proportion of retained pseudo-labeled samples (*Retention*). Ground-truth labels are used only for this post-hoc evaluation and are never involved in pseudo-label construction or model training.

As shown in Table 6, the proposed noise elimination generally improves the quality of the retained pseudo-labels. The pseudo-label error rate decreases on four of the five datasets, reducing from 27.4% to 23.3% on average. The effect is particularly pronounced on MUTAG, PROTEINS, and BAMultiShapes, where the error rate is reduced by 9.7, 5.9, and 3.9 percentage points, respectively.

More importantly, the elimination mechanism preferentially removes erroneous pseudo-labels rather than indiscriminately discarding training samples. On four datasets, the removal rate of incorrectly pseudo-labeled samples is higher than that of correctly pseudo-labeled samples. This difference is especially clear on BAMultiShapes and PROTEINS, where the corresponding removal-rate gaps reach 20.5 and 15.7 percentage points, respectively. Although the filtering effect is limited on NCI109, the overall results indicate that prototype deviation provides useful information for distinguishing unreliable pseudo-labels from reliable ones.

These observations complement the downstream ablation results from a pseudo-label-quality perspective. While removing pseudo-labeled samples inevitably trades supervision quantity for supervision quality, the proposed mechanism generally reduces the contamination of the retained supervision and therefore provides a cleaner training signal for subsequent optimization.

D.4 More Representation Visualization

Figure 4 provides additional visualization results under different observed positive ratios. At $r_{\text{obs}} = 0.1$, the two classes remain partially mixed in both stages, although the late-stage representation shows fewer scattered samples across class regions and a more compact overall structure. For $r_{\text{obs}} = 0.2$ and $r_{\text{obs}} = 0.3$, the late-stage representations exhibit clearer class separation and reduced overlap compared with their intermediate counterparts. In particular, positive and negative graphs are distributed in more distinct class-specific regions after further optimization, indicating

Table 7: Sensitivity analysis (%) of the graph-of-graphs neighborhood size k at $r_{\text{obs}} = 0.1$ across five datasets.

r_{obs}	k	MUTAG		NCI1		NCI109		PROTEINS		BAMultiShapes	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
0.1	2	28.5	33.3	36.7	45.8	30.9	37.0	42.5	62.3	52.2	73.7
	3	35.4	41.2	36.4	47.0	30.9	36.6	47.8	67.3	49.3	68.5
	4	29.5	35.3	37.4	48.4	31.8	38.5	43.8	63.0	49.0	71.1
	5	29.9	35.3	34.5	43.1	29.5	34.5	44.0	67.0	47.1	65.9

Table 8: Sensitivity analysis (%) of the augmentation ratio ρ at $r_{\text{obs}} = 0.1$ across five datasets.

r_{obs}	ρ	MUTAG		NCI1		NCI109		PROTEINS		BAMultiShapes	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
0.1	0.05	33.9	41.2	36.4	47.0	29.1	34.5	47.8	67.3	44.6	64.1
	0.1	35.4	41.2	37.3	48.4	31.8	38.5	49.3	71.3	49.3	68.5
	0.15	36.3	45.1	39.3	52.8	26.6	30.3	47.0	69.7	48.4	67.4
	0.2	31.3	37.3	37.3	48.7	30.5	36.1	50.4	74.3	44.2	60.4

that the learned representation space becomes increasingly aligned with the underlying class structure. The improvement is more evident when more verified positive graphs are available, as stronger supervision provides a more reliable basis for refining the representation geometry. These results are consistent with the expected effect of dynamically removing unreliable pseudo-labels during training, thereby reducing noisy supervision and promoting more discriminative graph representations.

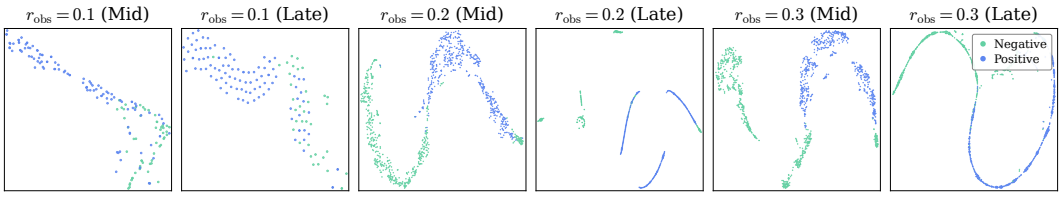


Figure 4: Visualization of graph representations at the intermediate and late training stages under different observed positive ratios r_{obs}. Colors denote the ground-truth graph classes, where green and blue represent negative and positive graphs, respectively.

D.5 MORE SENSITIVITY STUDY

The sensitivity studies in this section are conducted on a fixed validation split for hyperparameter analysis, rather than the ten-fold evaluation protocol used for the main experiments. Therefore, the absolute values reported here are not directly comparable with the main results, and the analysis focuses on the relative performance trends under different parameter settings.

As shown in Table 7, performance varies moderately with the graph-of-graphs neighborhood size k. A relatively small neighborhood generally performs well across datasets, while larger k provides no consistent improvement and may introduce less relevant relational information. Table 8 further examines the augmentation ratio ρ. Moderate perturbations maintain competitive performance, whereas stronger augmentation shows more dataset-dependent effects, suggesting that excessive perturbation may distort task-relevant graph structures. Overall, these results support relatively conservative settings for both neighborhood construction and graph augmentation.