POPGLAD: STRUCTURAL POPULATION SCORING FOR GRAPH-LEVEL ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph-level anomaly detection (GLAD) identifies graphs whose structure or attributes deviate from normal patterns. Many existing methods assess each graph through its departure from learned normality. Such scores can obscure a structural change that is weak in individual graphs but consistent across anomalies. To address this challenge, we propose PopGLAD, a population scoring method that leverages an unlabeled graph collection to capture shared evidence. For each graph, PopGLAD estimates a reference direction from the mean departure of the remaining unlabeled graphs and measures directional agreement relative to normal variation. PopGLAD evaluates this agreement in complementary structural and size views that capture local connectivity and graph scale, with normal statistics determining the comparison geometry and calibration. The final scores can be computed efficiently from shared population summaries without explicit pairwise graph comparisons. Our theoretical analysis bounds their pairwise ranking error and characterizes how the strength and stability of the population direction affect the reliability of scoring. Experiments on eight graph benchmarks demonstrate the superiority of PopGLAD over state-of-the-art GLAD baselines.

1 INTRODUCTION

Graph-level anomaly detection (GLAD) (Akoglu et al., 2015; Qiao et al., 2025) identifies graphs whose structural or attribute patterns differ from those of normal graphs. It has many real-world applications, such as molecular screening and brain network analysis, where detecting atypical compounds or abnormal connectivity can guide further investigation (Liu et al., 2023b). A graph may be anomalous because of a localized neighborhood pattern, its overall organization, or both (Ma et al., 2022). Since anomalous examples are often scarce, a common GLAD setting learns from normal graphs and detects on an unseen collection containing both normal and anomalous graphs (Qiu et al., 2022; Wang et al., 2025).

Existing GLAD methods typically encode node attributes and edges with graph neural networks (GNNs) (Kipf & Welling, 2017), then learn normality through a detection objective. For example, one-class classification based methods (Qiu et al., 2022) learn a compact region, e.g., a hypersphere, to enclose normal graph embeddings in the latent space, reconstruction methods (Kim et al., 2024) assess errors in recovering graph structure or attributes, and knowledge distillation (Ma et al., 2022) measures disagreement with a reference network at node and graph levels. These approaches translate the departure of a graph from learned normality into an anomaly score. Besides, self-interpretable methods (Liu et al., 2023b) also identify anomalies via explainable scoring for graph substructures. Recent methods also incorporate collection context by learning relations across graphs and views (Li et al., 2023) or through extracted subgraphs (Yang et al., 2026a).

However, an essential challenge is that meaningful structural differences do not necessarily lead to extreme individual scores. Reconstruction models can assign small errors to anomalous graphs (Kim et al., 2024), and class overlap in graph embedding spaces can hinder outlier ranking (Zhao & Akoglu, 2023). Consider a subtle local structural change that recurs across anomalous graphs, while normal graphs also vary substantially in other patterns. Individual scores can obscure such a shared change, whereas recurrence alone does not establish abnormality because normal graphs also share structural patterns. This motivates evaluating whether a graph's departure from normality aligns with a recurring *direction of change* across the unlabeled population, relative to normal variation.

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.

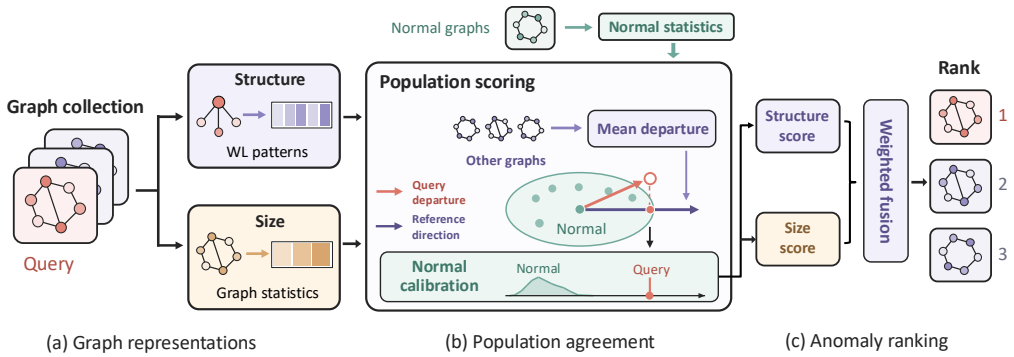

Figure 1: Overview of PopGLAD. (a) Structural and size representations. (b) Directional agreement with the mean departure of the remaining graphs, calibrated using normal variation. (c) Score fusion and anomaly ranking. All fitted statistics are estimated from normal graphs.

To capture this population-level evidence, we propose PopGLAD, a structural population scoring method for GLAD (Figure 1). For each graph, PopGLAD measures the agreement between its departure from normality and the average departure of the remaining graphs in the unlabeled collection. This average defines a reference direction that reinforces consistent changes while unrelated fluctuations tend to cancel. To operationalize this idea, sparse counts of attributed neighborhoods preserve local connectivity, while normal statistics weight structural patterns and calibrate directional agreement relative to normal variation. A complementary size view further retains graph-scale information, and the two views are fused using weights estimated from normal graphs. For efficient computation, shared population summaries enable scoring without storing pairwise graph similarities. We summarize our contributions as follows:

- We propose PopGLAD, a population scoring method for graph-level anomaly detection. It measures the agreement between the departure of a graph from normality and a reference direction estimated from the remaining graphs, while using normal variation to weight and calibrate complementary structural and size views.
- We derive a bound on pairwise ranking error for the final scores, accounting for finite reference uncertainty and queries entering one another's references. The analysis characterizes how the strength of the population departure and normal center error affect ranking reliability.
- Experiments on eight benchmarks demonstrate strong detection performance and computational efficiency. Further analyses validate the role of population agreement and characterize how reference quality influences anomaly detection performance.

2 Related Work

2.1 Graph-Level Anomaly Detection

Graph-level anomaly detection (GLAD) identifies graphs whose structure or attributes deviate from normal graph patterns (Akoglu et al., 2015; Qiao et al., 2025). The task requires capturing local irregularities and global structural differences while accounting for normal variation (Ma et al., 2022). Existing methods use different criteria to turn graph representations into anomaly evidence. One-class methods (Zhao & Akoglu, 2023; Qiu et al., 2022; Zhang et al., 2024) measure deviation from a normal region in graph representation space. Reconstruction and knowledge distillation (Luo et al., 2022; Ma et al., 2022) use errors in recovering graph information or matching reference representations. Moreover, MUSE (Kim et al., 2024) shows that multiple summaries of reconstruction errors retain information lost by their mean, illustrating how score aggregation affects detection. Methods for interpretable detection associate anomaly scores with explanatory substructures and prototypes (Liu et al., 2023b; Yang et al., 2025). Within the setting of training on normal graphs (Wang et al., 2025), PopGLAD assesses whether the departure of a graph from normality agrees with departures in the remaining unlabeled collection.

2.2 Structural Graph Representations

Structural representations determine which graph differences can contribute to anomaly scores. Graph kernels (Vishwanathan et al., 2010; Kriege et al., 2020) compare graphs through structural patterns, including propagated node information (Neumann et al., 2016) and counts of refined neighborhoods in Weisfeiler-Lehman kernels (Shervashidze et al., 2011). Neighborhood refinement also provides a basis for analyzing the expressive power of graph neural networks that use message passing (Xu et al., 2019; Morris et al., 2019). For explicit feature representations, hashing bounds the feature dimension (Weinberger et al., 2009), while frequency weighting reduces the relative influence of common patterns (Salton & Buckley, 1988). PopGLAD organizes attributed neighborhood counts into a shared sparse feature space, with frequency weights fitted from normal graphs. This representation lets recurring structural departures contribute to a population direction, while a separate size view retains information attenuated by structural normalization.

2.3 Population Information and Statistical Scoring

Collection information supports both models of sample interactions and statistical comparisons of distributions. NPT (Thimonier et al., 2024) uses training samples as context for reconstructing tabular inputs. For graphs, CVTGAD (Li et al., 2023) models relations across graphs and views, while GADTrans (Yang et al., 2026a) learns graph relationships through extracted subgraphs. At the distribution level, kernel mean embeddings (Sriperumbudur et al., 2010; Muandet et al., 2017) summarize collections in feature space, and witness functions (Gretton et al., 2012) evaluate samples against mean differences for distribution testing. Precision weighting incorporates feature covariance into this comparison (Kübler et al., 2022), with shrinkage stabilizing covariance estimation (Ledoit & Wolf, 2004; Chen et al., 2010). Compared with learned interaction models, PopGLAD represents the collection of evidence by the mean structural departure from normality. Relative to witness tests, it targets anomaly ranking within the unlabeled graph collection. Normal graphs determine how departures are centered, weighted, and calibrated, while the remaining unlabeled graphs provide the reference direction for each query.

3 Problem Formulation

Let $G = (V_G, E_G, X_G, B_G)$ be a nonempty simple undirected graph with node and edge attributes. We observe disjoint sets $\mathcal{F}$ and $\mathcal{C}$ of normal graphs for fitting and calibration, with $|\mathcal{F}| \geq 4$ and $|\mathcal{C}| \geq 2$, and a fixed unlabeled collection $\mathcal{U} = \{G_1, \ldots, G_n\}$ with $n \geq 2$. The goal is to assign larger scores to anomalous graphs in $\mathcal{U}$.

We consider transductive inference, where a query can use the other unlabeled graphs in $\mathcal{U}$. The collection is specified by the evaluation or deployment task. All fitted statistics and fusion weights depend on $\mathcal{F}$ and $\mathcal{C}$, and remain fixed during scoring. Each query is excluded from its own population reference. Computation batches partition the work on this fixed collection.

4 Methodology

PopGLAD detects graphs whose departures from normality agree with changes observed across the unlabeled collection. We first represent local structure and graph size, then use normal graphs to estimate feature weights and score scales. The remaining unlabeled graphs provide a reference direction for each query.

4.1 Graph Representations

A shared structural representation allows recurring graph changes to contribute to the same feature coordinates. Let $c_v^{(0)}$ encode the categorical attributes of node v, and let b_{uv} encode an edge attribute. We construct attributed Weisfeiler-Lehman neighborhoods (Shervashidze et al., 2011) through

$$c_v^{(\ell+1)} = H\left(c_v^{(\ell)}, |\mathcal{N}(v)|, \operatorname{sort}\{(b_{uv}, c_u^{(\ell)}) \mid u \in \mathcal{N}(v)\}\right). \tag{1}$$

Here H is a deterministic hash and the braces denote a multiset, which is empty for isolated nodes. Counts at depths $0, \ldots, L$ occupy separate blocks of width M, giving $h(G) \in \mathbb{R}^d$ with $d = (L+1)M$. For $N_f = |\mathcal{F}|$, let df_j count fitting graphs containing coordinate j. Define

$$w_j = 1 + \log \frac{N_f + 1}{\mathrm{df}_j + 1}, \qquad a(G) = w \odot \log(1 + h(G)), \qquad x_s(G) = \frac{a(G)}{\|a(G)\|_2}. \tag{2}$$

Frequency weighting emphasizes patterns uncommon in normal graphs. Logarithmic counts and normalization control feature magnitude while retaining neighborhood information subject to WL expressivity and hash collisions. Nonempty graphs have $\|a(G)\|_2 > 0$.

Normalization also reduces information about graph scale. We retain this information in a size view with inputs $\log(1 + |V_G|)$, $\log(1 + |E_G|)$, and the fraction of nonisolated nodes. Each undirected edge is counted once. An inverse hyperbolic sine transform followed by fitting mean and standard deviation scaling gives $x_a(G) \in \mathbb{R}^3$ (Appendix B). Both view transformations are fitted on $\mathcal{F}$ and fixed during scoring.

4.2 Normal Statistics Estimation

Normal variation provides the scale for comparing feature deviations. We use covariance shrinkage to stabilize the estimated precision weights and a separate normal calibration set to center and scale the directional scores. For each view $v \in \{s, a\}$, let $\widehat{\mu}_v$ be its fitting mean. In the structural view, let v_j and $h_j^{(4)}$ be the sample variance and empirical fourth central moment. On the active coordinates $J = \{j : \mathrm{df}_j > 0\}$, let $t > 0$ be the mean variance with a positive floor. For $\epsilon_D > 0$, set

$$b = \frac{1}{N_f} \sum_{j \in J} \max\left(h_j^{(4)} - \frac{N_f - 3}{N_f - 1} v_j^2, 0 \right), \quad D = \sum_{j \in J} (v_j - t)^2, \quad \alpha = \mathrm{clip}_{[0,1]} \frac{b}{\max(D, \epsilon_D)}. \tag{3}$$

The ratio compares estimated variance uncertainty with differences across coordinate variances. Following covariance shrinkage (Ledoit & Wolf, 2004), we use a base ridge $\gamma_0 > 0$ and define

$$\eta_s = (1 - \alpha)\gamma_0 + \alpha t, \qquad P_s = \mathrm{diag}\left((1 - \alpha)v_j + \eta_s\right)^{-1}. \tag{4}$$

The covariance interpolates between $\mathrm{diag}(v_j + \gamma_0)$ and tI, with eigenvalues at least $\min(\gamma_0, t)$. A small α preserves more of the empirical variance differences. A larger α moves the covariance toward the isotropic target. Appendix A.4 gives the moment calculation. The size precision P_a is the inverse, or pseudoinverse when singular, of a Ledoit-Wolf covariance estimate.

To measure normal variation along the weighted direction, estimate means c_v and sample covariances C_v using the same $K \geq 2$ selected calibration graphs in both views. Set $\Omega_v = C_v + \eta_v I$, with η_s from Equation (4) and $\eta_a > 0$ determined from fitting data. Calibration samples represent C_v implicitly. Fixed settings are given in Appendix C.

4.3 Population Scoring and Fusion

For query G_i, averaging the departures of the remaining graphs gives the reference displacement

$$\delta_{v,-i} = \frac{1}{n-1} \sum_{j \neq i} x_v(G_j) - \widehat{\mu}_v, \qquad s_{v,i} = \frac{(x_v(G_i) - c_v)^\top P_v \delta_{v,-i}}{\max\left(\sqrt{\delta_{v,-i}^\top P_v \Omega_v P_v \delta_{v,-i}}, \tau \right)}. \tag{5}$$

The numerator measures the query's agreement with this direction, weighted by normal precision. The fitting center $\widehat{\mu}_v$ defines the reference departure, while c_v centers its response on calibration graphs. Excluding the query makes the reference depend on departures in other graphs. The denominator scales agreement by normal variation along the same direction, with a fixed floor $\tau > 0$.

This score builds on mean witnesses and precision weighting (Gretton et al., 2012; Kübler et al., 2022). For a common reference, calibration preserves each view's ranking and gives zero empirical calibration mean and variance at most one (Appendices A.2 and A.3). When structural variance estimates require stronger shrinkage, we give the complementary size view more weight while keeping structure as the primary source of evidence. We combine the calibrated scores as

$$s_i = \theta_s s_{s,i} + \theta_a s_{a,i}, \qquad \lambda = \sqrt{\alpha}, \quad \theta_s = \frac{1}{1 + \lambda}, \quad \theta_a = \frac{\lambda}{1 + \lambda}. \tag{6}$$

The rule is estimated entirely from normal graphs. Appendix A.8 gives conditions for improving the ranking error bound, and Appendix D.1 further evaluates the chosen weights.

4.4 Theoretical Analysis

We first examine why directional agreement can detect anomalies with ordinary individual distances. In an isotropic Gaussian feature model with exact normal centering and a mean shift of fixed norm $r > 0$, the population direction has AUROC $\Phi(r/\sqrt{2})$, where Φ is the standard normal CDF. Squared distance has AUROC tending to $1/2$ as the number of unrelated dimensions grows. Appendix A.7 proves this comparison and quantifies the reference size needed to estimate the direction.

We next analyze the actual score with a finite reference, conditional on all fitted normal quantities. Put $A_v = \Omega_v^{1/2} P_v$, $y_v = A_v(x_v - \widehat{\mu}_v)$, $b_v = \Omega_v^{-1/2}(x_v - c_v)$, and $g_\tau(q) = q/\max(\|q\|_2, \tau)$. Then $s_i = \sum_v \theta_v b_v(G_i)^\top g_\tau(\bar{y}_{v,-i})$, where $\bar{y}_{v,-i}$ averages the remaining graphs.

For normal and anomaly distributions P_0, P_1 and reference mixture $Q_\pi = (1-\pi)P_0 + \pi P_1$, define $d_v = \mathbb{E}_{Q_\pi} y_v$, $\kappa_v = \|d_v\|_2$, $\rho_v = \max(\kappa_v, \tau)$, and $\sigma_v^2 = \mathbb{E}_{Q_\pi}\|y_v - d_v\|_2^2$. The score using the population direction is $s_\infty(G) = \sum_v \theta_v b_v(G)^\top g_\tau(d_v)$.

Theorem 1 (Ranking stability with a finite reference)**.** *Independently draw $G_1 \sim P_1$, $G_0 \sim P_0$, and $H_1, \ldots, H_k \sim Q_\pi$, with $k \geq 1$ and $\pi \in [0,1]$, and score the $n = k+2$ graphs with query exclusion. Assume $\|y_v\|_2 \leq B_v$ and $\|b_v\|_2 \leq R_v$ almost surely under both classes. For each view with $\theta_v > 0$ and $\xi_v \in (0,1)$, set*

$$u_v(k) = \min\left\{ \frac{\sigma_v}{\sqrt{k\xi_v}},\ \frac{\sigma_v + B_v\sqrt{2\log(1/\xi_v)}}{\sqrt{k}} \right\},$$
$$e_v(k) = \frac{k u_v(k) + 2B_v}{k+1}, \qquad \varepsilon_k = \sum_v \theta_v R_v \min\left\{ 2, \frac{e_v(k)}{\tau}, \frac{2e_v(k)}{\rho_v} \right\}. \tag{7}$$

For the ideal margin $T_\infty = s_\infty(G_1) - s_\infty(G_0)$,

$$\Pr\{s(G_1) \leq s(G_0)\} \leq \min\{1,\ \textstyle\sum_v \xi_v + \Pr(T_\infty \leq 2\varepsilon_k)\}. \tag{8}$$

The bound connects ranking reliability to the ideal class margin and reference uncertainty. The term $2B_v/(k+1)$ accounts for the other query entering each reference. The denominator floor keeps the result defined for weak and zero population directions. Appendix A.5 gives the proof, an explicit bound in terms of margin moments, and an extension under finite second moments.

Let $\Delta_v = \mathbb{E}_{P_1} x_v - \mathbb{E}_{P_0} x_v$ and $a_v = \widehat{\mu}_v - \mathbb{E}_{P_0} x_v$. The ideal mean margin is

$$m = \mathbb{E}T_\infty = \sum_v \theta_v \frac{\Delta_v^\top P_v(\pi\Delta_v - a_v)}{\rho_v}. \tag{9}$$

With exact centering and a population direction above the floor, prevalence cancels from this margin. The sampling term in relative reference error is $O(1/(\pi\sqrt{k}))$ at fixed confidence and geometry with uniformly bounded reference variance. Thus larger references improve direction estimation, while normal center accuracy and the aggregate anomaly shift determine its alignment with the classes (Appendix A.6).

4.5 Efficient Computation

The reference means differ only in the excluded query, so we reuse shared collection summaries to compute the scores. Let $z_{v,i} = x_v(G_i) - \widehat{\mu}_v$, $\bar{z}_v = n^{-1}\sum_i z_{v,i}$, and $\zeta_v = \widehat{\mu}_v - c_v$. Since $\delta_{v,-i} = (n\bar{z}_v - z_{v,i})/(n-1)$, the numerator equals

$$\frac{n(z_{v,i} + \zeta_v)^\top P_v \bar{z}_v - z_{v,i}^\top P_v z_{v,i} - \zeta_v^\top P_v z_{v,i}}{n-1}. \tag{10}$$

Shared calibration products similarly give the exact denominator (Appendix B). For S stored structural entries, dimension d, and K calibration graphs, computation after encoding costs $O(KS + Kd + nK)$, with $O(S + d + n)$ storage and $O(BK)$ additional workspace for batches of size B. These identities preserve query exclusion through shared summaries. Proposition 1 establishes invariance to node relabeling and consistent scores under collection permutations and computation batching.

Table 1: Anomaly detection performance (AUROC and AUPRC) on small and moderate benchmarks. We report mean ± standard deviation (%) over ten runs, and the best results are marked in bold.

Method	MUTAG		AIDS		DD		ER_MD	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GLocalKD	89.93 ± 0.89	95.44 ± 0.75	57.62 ± 2.84	51.05 ± 2.00	62.99 ± 2.08	86.03 ± 1.24	41.40 ± 9.80	42.10 ± 4.25
OCGTL	89.29 ± 0.53	95.43 ± 0.20	$\mathbf{99.67 \pm 0.02}$	$\mathbf{99.77 \pm 0.01}$	63.33 ± 3.73	80.89 ± 1.87	66.46 ± 0.39	57.74 ± 0.53
OCGIN	61.78 ± 17.22	78.16 ± 8.77	99.63 ± 0.05	99.74 ± 0.04	25.59 ± 1.02	66.02 ± 0.46	67.68 ± 1.42	58.90 ± 1.39
SIGNET	82.55 ± 4.61	90.93 ± 2.77	94.01 ± 0.73	93.07 ± 1.30	71.94 ± 1.88	88.74 ± 1.04	80.76 ± 1.56	76.69 ± 1.39
CVTGAD	35.10 ± 27.28	64.18 ± 13.36	99.51 ± 0.08	99.67 ± 0.06	29.83 ± 7.34	68.37 ± 3.73	68.28 ± 2.07	59.24 ± 1.57
GLADC	42.84 ± 5.25	68.55 ± 2.42	99.06 ± 0.09	98.32 ± 0.21	23.21 ± 2.88	64.75 ± 1.31	64.63 ± 1.35	58.33 ± 1.27
GOOD-D	78.39 ± 3.25	86.60 ± 2.06	92.16 ± 1.05	87.69 ± 2.79	41.39 ± 0.61	74.65 ± 0.53	73.29 ± 2.35	66.12 ± 4.73
GraphDE	27.01 ± 0.77	58.27 ± 0.31	62.96 ± 0.18	54.98 ± 0.09	57.64 ± 1.25	84.44 ± 0.60	51.11 ± 1.93	57.38 ± 4.13
MUSE	72.37 ± 5.13	85.99 ± 2.34	87.79 ± 13.46	82.36 ± 15.14	77.48 ± 1.40	91.99 ± 0.44	51.68 ± 3.98	46.95 ± 2.11
GUL	84.21 ± 2.88	91.76 ± 1.87	37.63 ± 5.80	40.04 ± 2.50	60.14 ± 1.45	83.90 ± 0.76	68.24 ± 0.76	59.75 ± 0.76
GLADPro	80.17 ± 7.69	89.81 ± 4.65	61.97 ± 13.28	54.80 ± 7.36	67.03 ± 6.51	87.13 ± 3.00	66.00 ± 6.81	58.85 ± 6.24
PopGLAD	$\mathbf{92.24 \pm 0.99}$	$\mathbf{95.97 \pm 0.54}$	99.45 ± 0.06	99.54 ± 0.06	$\mathbf{81.58 \pm 0.16}$	$\mathbf{93.41 \pm 0.04}$	$\mathbf{88.33 \pm 0.72}$	$\mathbf{89.95 \pm 0.55}$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. In this paper, we evaluate the proposed PopGLAD on eight graph benchmarks (Morris et al., 2020), including: (1) four small and moderate-sized datasets: MUTAG, AIDS, DD, ER_MD, and (2) four large-scale datasets: SW-620, MOLT-4, PC-3, and MCF-7. The detailed information of each dataset is illustrated in Appendix C.

Implementation details. Due to the length limitation of the paper, we provide the detailed implementation details in Appendix C.

Baseline methods. We compare PopGLAD with extensive state-of-the-art GLAD baselines, including: GLocalKD (Ma et al., 2022), OCGTL (Qiu et al., 2022), OCGIN (Zhao & Akoglu, 2023), SIGNET (Liu et al., 2023b), CVTGAD (Li et al., 2023), GLADC (Luo et al., 2022), GOOD-D (Liu et al., 2023a), GraphDE (Li et al., 2022), MUSE (Kim et al., 2024), GUL (Yang et al., 2026b), and GLADPro (Yang et al., 2025). We reproduce the baselines using public implementations with shared test partitions, and the settings reported in Appendix C.

Evaluation metrics. To evaluate the performance of each method on the anomaly detection task, we use two popular evaluation metrics across the experiment, including Area Under the ROC Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC) (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Particularly, we compute AUPRC using trapezoidal integration of the precision–recall curve. To make the results more reliable, we report the mean and standard deviation over ten runs.

5.2 ANOMALY DETECTION PERFORMANCE

To evaluate anomaly detection performance, we compare PopGLAD with eleven baselines on eight graph benchmarks in Tables 1 and 2. PopGLAD achieves the highest mean AUROC and AUPRC on seven benchmarks. For example, on the moderate-size dataset DD, PopGLAD exceeds the strongest baseline, MUSE, by 4.10% in AUROC and 1.42% in AUPRC. AIDS is the only exception, where PopGLAD and the strongest baseline, OCGTL, already exceed 99% on both metrics.

The improvement also extends to large-scale graph datasets with competitive baselines. For example, on MOLT-4, PopGLAD exceeds the strongest baseline, GOOD-D, by 19.20 percentage points in AUROC and 32.43 points in AUPRC. Improvements in both metrics hold across all four larger graph datasets, with margins that are large relative to the reported variation across runs. Across the two groups of benchmarks, PopGLAD provides its largest improvements where baseline scores offer limited separation between normal and anomalous graphs, while remaining competitive when detection performance is close to saturation. The experimental results across these settings support the value of population evidence across different graph collections.

Table 2: Anomaly detection performance (AUROC and AUPRC) on large-scale graph benchmarks.

Method	SW-620		MOLT-4		PC-3		MCF-7	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GLocalKD	49.56 ± 2.35	24.40 ± 1.06	51.16 ± 2.05	31.52 ± 1.29	49.16 ± 2.15	24.14 ± 1.06	48.73 ± 1.94	31.06 ± 1.47
OCGTL	66.57 ± 0.03	40.28 ± 0.06	64.47 ± 0.05	43.92 ± 0.06	68.32 ± 0.04	41.41 ± 0.12	65.13 ± 0.02	47.64 ± 0.03
OCGIN	66.49 ± 0.07	40.18 ± 0.11	64.55 ± 0.11	43.83 ± 0.17	68.37 ± 0.14	41.37 ± 0.17	65.24 ± 0.20	47.60 ± 0.35
SIGNET	67.10 ± 1.50	36.60 ± 0.64	65.63 ± 1.65	42.19 ± 0.85	67.88 ± 1.62	37.58 ± 1.30	65.32 ± 2.63	44.37 ± 1.90
CVTGAD	53.98 ± 0.55	26.34 ± 0.68	63.46 ± 0.24	43.72 ± 0.36	55.00 ± 1.01	26.37 ± 0.62	63.99 ± 0.33	45.40 ± 0.66
GLADC	60.42 ± 1.76	36.97 ± 1.58	60.51 ± 0.91	41.84 ± 0.97	67.39 ± 0.65	41.53 ± 0.42	64.55 ± 0.49	48.19 ± 0.39
GOOD-D	65.74 ± 1.23	39.06 ± 1.42	66.24 ± 0.74	45.76 ± 0.90	67.48 ± 1.13	40.49 ± 1.25	64.11 ± 1.00	46.20 ± 1.17
GraphDE	55.05 ± 0.03	24.59 ± 0.01	53.45 ± 0.02	29.98 ± 0.01	54.71 ± 0.03	23.30 ± 0.03	54.16 ± 0.04	31.33 ± 0.02
MUSE	61.09 ± 3.09	31.93 ± 4.27	60.57 ± 1.25	37.03 ± 1.58	61.22 ± 1.99	29.92 ± 2.96	60.88 ± 2.11	38.06 ± 2.55
GUL	43.38 ± 0.86	21.52 ± 0.28	43.87 ± 0.56	27.64 ± 0.21	44.16 ± 2.14	20.11 ± 0.57	45.05 ± 1.64	27.41 ± 0.72
GLADPro	50.68 ± 5.88	23.99 ± 3.16	51.90 ± 5.16	30.70 ± 2.81	51.46 ± 6.90	25.05 ± 4.46	52.41 ± 4.30	32.24 ± 3.74
PopGLAD	$\mathbf{89.24 \pm 0.10}$	$\mathbf{79.35 \pm 0.09}$	$\mathbf{85.44 \pm 0.17}$	$\mathbf{78.19 \pm 0.16}$	$\mathbf{88.87 \pm 0.11}$	$\mathbf{78.77 \pm 0.18}$	$\mathbf{86.76 \pm 0.08}$	$\mathbf{80.79 \pm 0.12}$

Table 3: Ablation study on four large-scale graph datasets. Note that all variants share graph features and normal partitions. Best results are marked in bold.

Method	SW-620		MOLT-4		PC-3		MCF-7	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Individual distance	55.04 ± 0.16	26.42 ± 0.08	55.32 ± 0.09	33.24 ± 0.14	55.40 ± 0.21	26.27 ± 0.14	56.49 ± 0.20	34.49 ± 0.13
Mean witness	83.34 ± 0.21	67.27 ± 0.43	79.87 ± 0.17	68.41 ± 0.35	83.97 ± 0.12	69.52 ± 0.27	81.07 ± 0.18	70.41 ± 0.35
Precision witness	86.01 ± 0.08	75.20 ± 0.06	83.31 ± 0.12	73.18 ± 0.12	85.46 ± 0.12	74.73 ± 0.19	82.94 ± 0.07	77.64 ± 0.11
PopGLAD	$\mathbf{89.24 \pm 0.10}$	$\mathbf{79.35 \pm 0.09}$	$\mathbf{85.44 \pm 0.17}$	$\mathbf{78.19 \pm 0.16}$	$\mathbf{88.87 \pm 0.11}$	$\mathbf{78.77 \pm 0.18}$	$\mathbf{86.76 \pm 0.08}$	$\mathbf{80.79 \pm 0.12}$

5.3 Ablation Study

To assess how population information contributes to anomaly detection, we conduct an ablation study to compare three scoring variants built from the same structural features and normal partitions, including: (1) ***Individual distance***: measures each graph's squared distance from the normal center using the estimated structural precision. (2) ***Mean witness***: measures the inner product between the deviation of the graph and the average deviation of the remaining test graphs, and (3) ***Precision witness***: further weights this inner product using a diagonal precision matrix computed from normal feature variances with a fixed ridge.

Table 3 shows that population agreement improves anomaly detection performance on all four graph datasets. On SW-620, the mean witness improves AUROC over individual distance by 28.30%. This gain indicates that agreement with the reference direction provides useful information beyond the individual distance score. Precision weighting further improves AUPRC over the mean witness by 7.93% on SW-620. This gain supports evaluating shared deviations relative to normal variation, since agreement along features that are stable among normal graphs provides stronger anomaly evidence. PopGLAD achieves the best performance on all four graph datasets. Compared with the precision witness, it further improves AUROC by 3.23% and AUPRC by 4.15% on SW-620. The largest improvement occurs when replacing individual distance with population agreement, while precision weighting and the full scoring rule provide consistent further gains.

5.4 Effect of the Reference Population

To examine how reference size affects detection, we vary the number of reference graphs while keeping the fitted model and a disjoint query set fixed. Figure 2(a,b) shows that larger references improve both metrics on all four graph datasets. On SW-620, the complete reference pool gives 88.09% AUROC, exceeding the result with 128 reference graphs by 17.78%. The gains support averaging over more reference graphs to reduce sampling variation in the direction used to evaluate each query.

To separate the effect of reference composition from its size, we fix the reference at 1024 graphs and vary anomaly prevalence in Figure 2(c,d). At 1% anomalies, AUROC remains below 57% on every graph dataset and improves as prevalence increases. With exact normal centering, prevalence cancels from the ideal score when population directions exceed the floor in all active views. In a finite

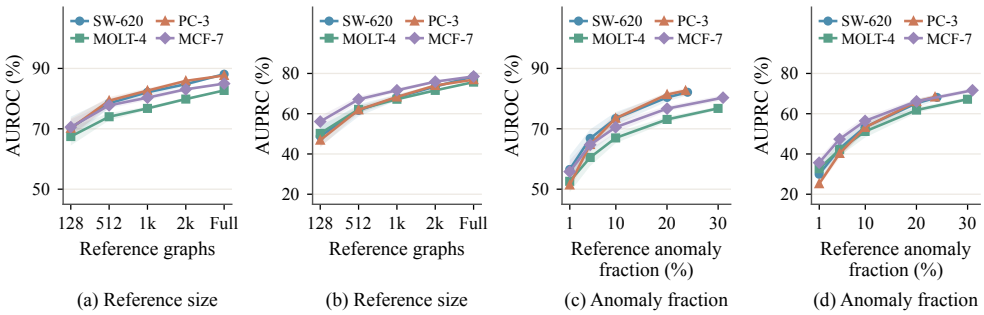

Figure 2: Effect of reference size (a,b) and anomaly fraction at a fixed reference size of 1024 in (c,d). Queries are fixed and disjoint from references. Lines and bands show the mean ± standard deviation across ten resamples. "Full" denotes the complete reference pool.

reference, a weaker mean deviation is more affected by sampling noise. Detection, therefore, depends on the strength of the aggregate anomaly deviation relative to estimation error, as characterized by Theorem 1.

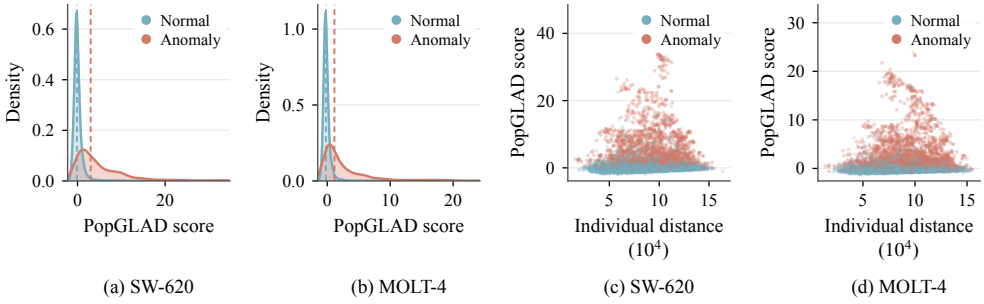

Figure 3: Score analysis on SW-620 and MOLT-4. (a,b) Class conditional densities with dashed medians. (c,d) Individual distance versus PopGLAD score for all queries in one run per dataset.

5.5 Score Analysis and Structural Sensitivity

To understand which anomalies benefit from population comparison, we examine score distributions and individual distances in Figure 3. Normal scores concentrate near zero, while anomalies have higher medians and longer positive tails. In the visualized runs for SW-620 and MOLT-4, more than 90% of anomalies fall below the 95th percentile of normal individual distances on each dataset. A threshold at this normal percentile would thus miss most anomalies. Many of these same graphs receive high PopGLAD scores. Together with the witness controls in Table 3, this separation shows how shared deviation directions provide anomaly evidence among graphs with comparable individual distances.

To assess sensitivity to connectivity changes, we apply edge edits that preserve graph size and node degrees. Figure 4 reports feature and score changes for 100 graph pairs per dataset. At least 90% of edited pairs have different representations, while controls with identical neighborhood representations retain identical features and scores. These edits leave the size view unchanged, so the score differences reflect structural information. Their varying magnitudes follow the scoring geometry: under the shared external reference, the effect of a feature change depends on its projection onto the reference direction after precision weighting.

5.6 Computational Efficiency

To evaluate the computational cost of population scoring, we compare total runtime with seven baselines on the four larger graph datasets in Table 4. Timings include data loading, graph preparation, fitting, and scoring under the same setting described in Appendix C.

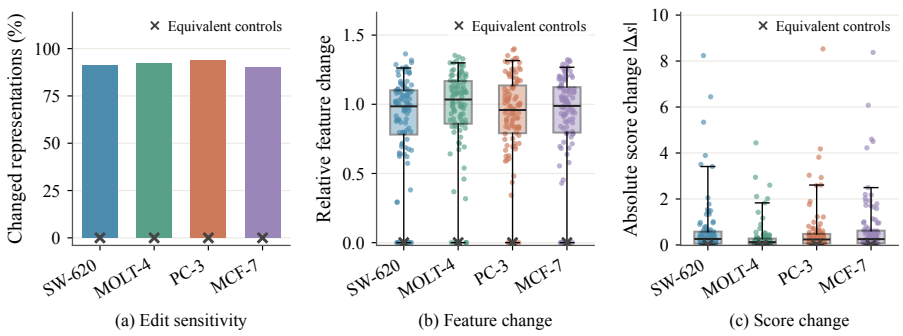

Figure 4: Effects of degree preserving edits on representations and scores (100 graph pairs per dataset). Dots show graph pairs, and whiskers mark the 5th and 95th percentiles. Crosses indicate controls with identical neighborhood representations.

PopGLAD has the shortest runtime on all four graph datasets. Relative to OCGIN, the fastest baseline on MOLT-4, it reduces total runtime by 67.41%. Direct estimation of normal statistics limits fitting cost, and shared collection summaries reuse the reference computations across queries. These summaries give the exact scores with query exclusion while avoiding a matrix of pairwise graph similarities, as described in Section 4.5. Together, Tables 2 and 4 show improved detection performance at a lower measured computational cost.

Table 4: Mean total runtime (s) with settings given in Appendix C. Bold indicates the fastest method.

Method	SW-620	MOLT-4	PC-3	MCF-7
GLocalKD	177.78	149.45	104.03	133.50
OCGTL	62.51	58.62	42.17	44.10
OCGIN	52.86	54.61	34.29	37.45
SIGNET	87.60	74.98	87.01	57.54
MUSE	253.74	290.81	190.87	147.83
GUL	112.94	75.68	76.70	64.31
GLADPro	74.35	63.43	47.15	66.76
PopGLAD	**17.80**	**17.80**	**12.39**	**12.24**

5.7 Additional Experimental Analysis

We provide more experimental analysis in the appendix. Specifically, Appendix D.1 examines covariance estimation and view fusion, and Appendix D.2 studies the stability of the estimated reference direction. Appendix D.3 extends the score distributions and distance comparisons to PC-3 and MCF-7. Appendix D.4 examines cancellation between opposing anomaly shifts. Appendices D.5 and D.6 examine sensitivity to calibration size and numerical consistency across query orders and computation batch sizes. Appendix D.7 evaluates anomaly recall under limited inspection budgets.

6 Conclusion

In this paper, we propose PopGLAD, which detects graph anomalies by comparing the departure of each graph from normality with the mean departure of the remaining unlabeled graphs. Normal statistics weight and calibrate structural and size evidence, while shared summaries enable efficient scoring. Our theoretical analysis bounds ranking error and relates it to the ideal class margin, reference uncertainty, and normal center accuracy. Experiments show strong detection performance across eight benchmarks and lower runtime compared to the latest GLAD baselines. These findings highlight that consistent structural departures across graphs can provide useful anomaly evidence when individual distances offer limited separation. However, detection can be limited by rare anomalies in finite references or by cancellation between opposing shifts. Future work will explore improved reference estimation for rare anomalies and multiple reference directions for heterogeneous anomaly patterns.

AI USE STATEMENT

We used generative AI tools to improve manuscript clarity, assist with literature discovery, and provide feedback on the presentation and interpretation of experimental results. The authors are responsible for reviewing and verifying all content prepared with AI assistance and take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses existing public graph benchmarks. Anomaly scores guide further inspection of graphs. For molecular datasets, conclusions about biological activity, safety, or clinical utility require domain validation. Since population scores depend on the reference collection used at inference, applications should use a reference population appropriate to their intended context.

REPRODUCIBILITY STATEMENT

Section 4 specifies the graph representations, fitted quantities, and scoring rule. Appendix A provides the theoretical assumptions and complete proofs, and Appendix B describes the algorithm and sparse computation. Appendix C specifies dataset splits, feature preprocessing, hyperparameters, and evaluation protocols.

REFERENCES

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688, 2015.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the International Conference on Machine Learning*, pp. 233–240, 2006.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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.

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

Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):6, 2020.

Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pp. 1403–1419. PMLR, 2022.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

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

Zenan Li, Qitian Wu, Fan Nie, and Junchi Yan. Graphde: A generative framework for debiased learning and out-of-distribution detection on graphs. *Advances in Neural Information Processing Systems*, 35:30277–30290, 2022.

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*, pp. 339–347, 2023a.

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, 2023b.

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.

Rongrong Ma, Guansong Pang, Ling Chen, and Anton Van Den Hengel. Deep graph-level anomaly detection by glocal knowledge distillation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 704–714, 2022.

Colin McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4602–4609, 2019.

Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. Tudataset: A collection of benchmark datasets for learning with graphs. In *ICML Workshop on Graph Representation Learning and Beyond*, 2020.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

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

Hezhe Qiao, Hanghang Tong, Bo An, Irwin King, Charu Aggarwal, and Guansong Pang. Deep graph anomaly detection: A survey and new perspectives. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5106–5126, 2025.

Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. Raising the bar in graph-level anomaly detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2196–2203. Proceedings of the International Joint Conferences on Artificial Intelligence Organization, 2022.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. Beyond individual input for deep anomaly detection on tabular data. In *Proceedings of the International Conference on Machine Learning*, pp. 48097–48123, 2024.

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.

Yili Wang, Yixin Liu, Xu Shen, Chenyu Li, Rui Miao, Kaize Ding, Ying Wang, Shirui Pan, and Xin Wang. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark. In *Proceedings of the International Conference on Learning Representations*, pp. 91903–91929, 2025.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1113–1120, 2009.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *Proceedings of the International Conference on Learning Representations*, 2019.

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*, pp. 3586–3597, 2025.

Zhenyu Yang, Ge Zhang, Jia Wu, Jian Yang, Hao Peng, and Pietro Liò. Learning from graph-graph relationship: a new perspective on graph-level anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 38(1):428–441, 2026a.

Zhenyu Yang, Ge Zhang, Shan Xue, Xiaoxiao Ma, Jian Yang, Hao Peng, Amin Beheshti, and Jia Wu. Revisiting graph-level anomaly detection: From partially to fully unsupervised learning. In *Proceedings of the ACM Web Conference 2026*, pp. 901–912, 2026b.

Yunhe Zhang, Yan Sun, Jinyu Cai, and Jicong Fan. Deep orthogonal hypersphere compression for anomaly detection. In *Proceedings of the International Conference on Learning Representations*, volume 2024, pp. 18066–18090, 2024.

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 Theoretical Analysis and Proofs

A.1 Notation and Scoring Representation

All scoring and ranking statements condition on the normal fitting and calibration data. The feature maps, fitting centers $\widehat{\mu}_v$, calibration centers c_v, precision matrices P_v, calibration matrices Ω_v, and fusion weights θ_v are therefore fixed. Each P_v is symmetric positive semidefinite, each Ω_v is positive definite, and $\tau > 0$. The structural precision is positive definite. Sums over views include those with $\theta_v > 0$, and the weights sum to one. The selected calibration subset is denoted by $\mathcal{C}_K \subseteq \mathcal{C}$, with $K \geq 2$.

Queries and reference graphs are sampled independently of the normal fitting and calibration data. Let P_0, P_1 be the normal and anomaly distributions, and let $Q_\pi = (1-\pi)P_0 + \pi P_1$ with $\pi \in [0,1]$. Define

$$A_v = \Omega_v^{1/2} P_v, \qquad y_v(G) = A_v(x_v(G) - \widehat{\mu}_v), \qquad b_v(G) = \Omega_v^{-1/2}(x_v(G) - c_v). \tag{11}$$

For $g_\tau(q) = q/\max(\|q\|_2, \tau)$, the implemented score is

$$s(G_i) = \sum_v \theta_v b_v(G_i)^\top g_\tau\left(\frac{1}{n-1}\sum_{j\neq i} y_v(G_j)\right). \tag{12}$$

Indeed, $b_v(G_i)^\top A_v \delta_{v,-i} = (x_v(G_i) - c_v)^\top P_v \delta_{v,-i}$ and $\|A_v\delta_{v,-i}\|_2^2 = \delta_{v,-i}^\top P_v \Omega_v P_v \delta_{v,-i}$. This form expresses the full score as a projection onto a normalized reference direction, including its denominator floor and query exclusion.

A.2 Mean Witnesses and Query Exclusion

The numerator has a direct connection to mean witnesses (Gretton et al., 2012; Kübler et al., 2022). For one view, define the positive semidefinite kernel $k_P(x, x') = x^\top P x'$. Let $\widehat{Q}$ be a reference distribution with mean $\bar{x}_{\widehat{Q}}$ and let $\widehat{P}_{\mathcal{F}}$ be the empirical fitting distribution. Their mean witness is

$$\begin{aligned} f_{\widehat{Q},\mathcal{F}}(x) &= \mathbb{E}_{X'\sim\widehat{Q}} k_P(x, X') - \mathbb{E}_{X'\sim\widehat{P}_{\mathcal{F}}} k_P(x, X') \\ &= x^\top P(\bar{x}_{\widehat{Q}} - \widehat{\mu}). \end{aligned} \tag{13}$$

Taking $\widehat{Q}$ as the empirical distribution of the remaining unlabeled graphs gives the numerator $f_{\widehat{Q},\mathcal{F}}(x_i) - f_{\widehat{Q},\mathcal{F}}(c)$. For a common reference, calibration applies a common offset and positive scale, preserving the ranking within a view. PopGLAD evaluates a separate reference for each query and calibrates the views before fusion.

To see the effect of query exclusion, put $z_i = x_i - \widehat{\mu}$ and $\bar{z} = n^{-1}\sum_i z_i$. When $c = \widehat{\mu}$,

$$z_i^\top P \delta_{-i} = \frac{n}{n-1} z_i^\top P \bar{z} - \frac{1}{n-1} z_i^\top P z_i. \tag{14}$$

The subtraction removes the nonnegative contribution from the query's own squared distance. For general c, expanding the numerator gives Equation (10). Query exclusion also changes the directional variance in the denominator. Under independent sampling, each query is independent of its own reference. The two final scores remain dependent because each query enters the other's reference. Theorem 1 accounts for this dependence.

A.3 Invariance, Continuity, and Calibration

Proposition 1 (Invariance and continuous shrinkage)**.** *PopGLAD is invariant to node relabeling that preserves attributes. Its score vector is equivariant to permutations of a fixed unlabeled collection and invariant to the partition of that collection into computation batches. For fixed transformed features, fitted moments, base ridge, and shrinkage target, the final scores are continuous in* $\alpha \in [0,1]$.

Proof. Node relabeling preserves the initial attributes and every neighborhood multiset in Equation (1). Induction on the refinement depth preserves all counted labels. Deterministic hashing

preserves their coordinates, and the size features also remain unchanged. Normal fitting and calibration therefore produce the same quantities. Permuting the unlabeled collection preserves its feature sum. Subtracting the corresponding query gives the same reference and score for each graph. Computation batches partition the evaluation of these fixed expressions.

Each structural covariance entry satisfies

$$d_j(\alpha) = (1-\alpha)(v_j + \gamma_0) + \alpha t \geq \min(\gamma_0, t) > 0. \tag{15}$$

Its inverse and $\eta_s(\alpha)$ are continuous on $[0,1]$. Thus P_s and $\Omega_s = C_s + \eta_s I$ vary continuously. The score denominator is at least τ, and the fusion coefficients are continuous functions of $\sqrt{\alpha}$. Composition establishes continuity of the final scores. The permutation and batching identities hold in exact arithmetic. Appendix D.6 evaluates numerical agreement. □

For a common reference δ_v, let $s_v(G, \delta_v)$ be the score in Equation (5) evaluated with that direction for every $G \in \mathcal{C}_K$. The empirical calibration mean is zero. The sample variance is

$$\widehat{\mathrm{Var}}_{\mathcal{C}_K}(s_v) = \frac{\delta_v^\top P_v C_v P_v \delta_v}{\max\{\delta_v^\top P_v \Omega_v P_v \delta_v, \tau^2\}} \leq 1. \tag{16}$$

The inequality follows from $\Omega_v = C_v + \eta_v I$ with $\eta_v > 0$. Applying the triangle inequality to the centered calibration score vectors gives

$$\widehat{\mathrm{Var}}_{\mathcal{C}_K}\left(\sum_v \theta_v s_v\right) \leq \left(\sum_v \theta_v \sqrt{\widehat{\mathrm{Var}}_{\mathcal{C}_K}(s_v)}\right)^2 \leq 1. \tag{17}$$

These empirical identities hold for the selected calibration graphs under common reference directions and permit arbitrary dependence between views. They provide the scale used for fusion. The ranking results below address independently drawn queries.

The same score definition relates structural changes to their directional effect. For two graphs evaluated against a common external reference,

$$s_v(G, \delta_v) - s_v(G', \delta_v) = \frac{(x_v(G) - x_v(G'))^\top P_v \delta_v}{\max\{\sqrt{\delta_v^\top P_v \Omega_v P_v \delta_v}, \tau\}}. \tag{18}$$

Identical representations give identical scores. A retained feature change affects the score through its projection onto the weighted reference direction. Edits that preserve graph size and all node degrees leave the size view unchanged, which connects this identity to the structural analysis in Section 5.5.

A.4 Variance Estimation and Precision Stability

The structural moments are

$$v_j = \frac{1}{N_f - 1} \sum_{G \in \mathcal{F}} (x_{s,j}(G) - \widehat{\mu}_{s,j})^2, \qquad h_j^{(4)} = \frac{1}{N_f} \sum_{G \in \mathcal{F}} (x_{s,j}(G) - \widehat{\mu}_{s,j})^4. \tag{19}$$

Coordinates absent from $\mathcal{F}$ have both moments zero. The shrinkage ratio uses the active set J, and the positive ridge gives every coordinate a finite precision.

For a fixed scalar feature with variance ν and fourth central moment μ_4, the sample variance $\widehat{\nu}$ from $N \geq 2$ independent and identically distributed observations satisfies

$$\mathrm{Var}(\widehat{\nu}) = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\nu^2\right). \tag{20}$$

To verify the identity, center the observations as Y_i and put $S_1 = \sum_i Y_i$ and $S_2 = \sum_i Y_i^2$. Independence and identical distributions give

$$\mathbb{E}[S_2^2] = \mathbb{E}[S_2 S_1^2] = N\mu_4 + N(N-1)\nu^2, \qquad \mathbb{E}[S_1^4] = N\mu_4 + 3N(N-1)\nu^2.$$

Expanding $\widehat{\nu} = (S_2 - S_1^2/N)/(N-1)$ yields

$$\mathbb{E}[\widehat{\nu}^2] = \frac{\mu_4}{N} + \frac{N^2 - 2N + 3}{N(N-1)}\nu^2.$$

Subtracting $(\mathbb{E}\widehat{\nu})^2 = \nu^2$ proves Equation (20).

Substituting fitted moments and taking positive parts gives the uncertainty estimate b in Equation (3). The term D measures dispersion around the isotropic target. The variance identity applies to a fixed feature map with independent observations. PopGLAD uses it as a moment approximation for the representation whose frequency weights are also fitted on $\mathcal{F}$. Together with the estimated target, this yields the regularization rule used in the method.

For fixed α, γ_0, t, two nonnegative variance vectors $\mathbf{v}, \widetilde{\mathbf{v}}$ satisfy

$$\|P_s(\mathbf{v}) - P_s(\widetilde{\mathbf{v}})\|_{\mathrm{op}} \le \frac{1-\alpha}{\eta_s^2}\|\mathbf{v} - \widetilde{\mathbf{v}}\|_\infty. \tag{21}$$

At coordinate j, the absolute difference is

$$\frac{(1-\alpha)|v_j - \widetilde{v}_j|}{((1-\alpha)v_j + \eta_s)((1-\alpha)\widetilde{v}_j + \eta_s)}.$$

Both denominator factors are at least η_s. Taking the maximum proves the bound and shows how regularization controls sensitivity to coordinate variance errors at a fixed shrinkage level and target.

A.5 Ranking Stability with a Finite Reference

Let

$$d_v = \mathbb{E}_{Q_\pi} y_v, \qquad \kappa_v = \|d_v\|_2, \qquad \rho_v = \max(\kappa_v, \tau), \qquad \sigma_v^2 = \mathbb{E}_{Q_\pi}\|y_v - d_v\|_2^2. \tag{22}$$

Theorem 1 assumes $\|y_v(G)\|_2 \le B_v$ and $\|b_v(G)\|_2 \le R_v$ almost surely under both classes. Since structural features have unit norm, valid structural bounds are $B_s = 2\|A_s\|_{\mathrm{op}}$ and $R_s = 2\|\Omega_s^{-1/2}\|_{\mathrm{op}}$. The same assumption holds for the size view when its transformed support is bounded. Proposition 2 extends the analysis to finite second moments.

The sampling model draws one query from each class and k independent mixture references, all mutually independent. For $0 < \pi < 1$, this also describes two designated positions in an independent mixture collection conditional on their opposite labels. The probability in Theorem 1 is the ranking error for this sampled pair.

Stability of the reference direction. For any vectors u, d, including zero,

$$\|g_\tau(u) - g_\tau(d)\|_2 \le \min\left\{2, \frac{\|u-d\|_2}{\tau}, \frac{2\|u-d\|_2}{\max(\|d\|_2, \tau)}\right\}. \tag{23}$$

Both outputs lie in the unit ball, giving the bound by two. The map $g_\tau(u)$ is the Euclidean projection of u/τ onto that ball. If p, p' are the projections of a, a', their projection inequalities give

$$\|p - p'\|_2^2 \le (p-p')^\top(a-a') \le \|p-p'\|_2\|a-a'\|_2.$$

This proves the bound by $\|u-d\|_2/\tau$. For the final bound, put $a = \max(\|u\|_2, \tau)$ and $b = \max(\|d\|_2, \tau)$. Since $|a-b| \le \|u-d\|_2$ and $\|u\|_2/a \le 1$,

$$\left\|\frac{u}{a} - \frac{d}{b}\right\|_2 \le \frac{\|u-d\|_2}{b} + \frac{\|u\|_2}{a}\frac{|a-b|}{b} \le \frac{2\|u-d\|_2}{b}.$$

The three bounds together cover directions both above and below the denominator floor.

Reference concentration. For independent $H_1, \ldots, H_k \sim Q_\pi$, put $\bar{y}_v = k^{-1}\sum_j y_v(H_j)$. Then

$$\mathbb{E}\|\bar{y}_v - d_v\|_2^2 = \frac{\sigma_v^2}{k}, \qquad \mathbb{E}\|\bar{y}_v - d_v\|_2 \le \frac{\sigma_v}{\sqrt{k}}.$$

Markov's inequality gives radius $\sigma_v/\sqrt{k\xi_v}$ with failure probability at most ξ_v. Changing one reference changes the norm $\|\bar{y}_v - d_v\|_2$ by at most $2B_v/k$. For $B_v > 0$, the bounded differences inequality (McDiarmid, 1989) therefore gives

$$\Pr\left\{\|\bar{y}_v - d_v\|_2 > \frac{\sigma_v}{\sqrt{k}} + t\right\} \le \exp\left(-\frac{kt^2}{2B_v^2}\right), \qquad t > 0. \tag{24}$$

This gives the second radius in $u_v(k)$. Selecting the smaller of the two deterministic radii retains failure probability at most ξ_v. When $B_v = 0$, the reference and its error are identically zero.

Proof of Theorem 1. Apply the concentration bound in every active view. A union bound gives an event $\mathcal{E}$ of probability at least $1 - \sum_v \xi_v$ on which $\|\bar{y}_v - d_v\|_2 \le u_v(k)$ for all views. The query references are

$$q_{v,1} = \frac{k\bar{y}_v + y_v(G_0)}{k+1}, \qquad q_{v,0} = \frac{k\bar{y}_v + y_v(G_1)}{k+1}. \tag{25}$$

Since $\|d_v\|_2 \le B_v$, on $\mathcal{E}$ each reference satisfies

$$\|q_{v,c} - d_v\|_2 \le \frac{ku_v(k) + \|y_v(G_{1-c}) - d_v\|_2}{k+1} \le e_v(k), \qquad c \in \{0,1\}.$$

The term $2B_v/(k+1)$ accounts for the query of the other class entering the reference. By Equation (23) and the Cauchy-Schwarz inequality,

$$|s(G_c) - s_\infty(G_c)| \le \sum_v \theta_v R_v \min\left\{2, \frac{e_v(k)}{\tau}, \frac{2e_v(k)}{\rho_v}\right\} = \varepsilon_k. \tag{26}$$

Consequently, an incorrect or tied ranking on $\mathcal{E}$ implies $T_\infty \le 2\varepsilon_k$. Taking the union of this margin event and $\mathcal{E}^c$ proves Equation (8). The argument permits dependence between views and between the two final scores. □

A bound using margin moments. Let $m = \mathbb{E}T_\infty$ and $V = \mathrm{Var}(T_\infty)$. If $m > 2\varepsilon_k$, Cantelli's inequality gives

$$\Pr\{s(G_1) \le s(G_0)\} \le \min\left\{1,\ \sum_v \xi_v + \frac{V}{V + (m - 2\varepsilon_k)^2}\right\}. \tag{27}$$

Counting ties as errors also gives an upper bound on $1 - \mathrm{AUROC}$ under the usual convention of assigning half credit to ties. The expression identifies favorable conditions as a large ideal mean margin, small margin variance, and an accurately estimated reference direction.

Ranking convergence and external references. Let $T_k = s(G_1) - s(G_0)$. On $\mathcal{E}$, $|T_k - T_\infty| \le 2\varepsilon_k$, so

$$|\Pr(T_k \le 0) - \Pr(T_\infty \le 0)| \le \sum_v \xi_v + \Pr(|T_\infty| \le 2\varepsilon_k). \tag{28}$$

For fixed fitted quantities and $\tau > 0$, choose $\xi_v = k^{-2}$ for $k \ge 2$. The radius from bounded differences tends to zero, so $\varepsilon_k \to 0$. If $\Pr(T_\infty = 0) = 0$, the pairwise error converges to the error under the population direction.

For k independent external references disjoint from both queries, each query uses $\bar{y}_v$. The same proof applies with $e_v(k) = u_v(k)$. This isolates reference sampling error from the additional effect of queries entering each other's references.

An extension under finite second moments. The following result covers transformed features with unbounded support and finite second moments.

Proposition 2 (Ranking under finite second moments)**.** *Under the same independent sampling model and fixed fitted quantities, assume finite class second moments of y_v and b_v. Define, for $c \in \{0,1\}$,*

$$\beta_{cv}^2 = \mathbb{E}_{P_c}\|b_v(G)\|_2^2, \qquad \omega_{cv} = \mathbb{E}_{P_c}\|y_v(G) - d_v\|_2^2, \qquad L_v = \min\{\tau^{-1}, 2\rho_v^{-1}\}. \tag{29}$$

Set

$$\chi_c(k) = \sum_v \theta_v \beta_{cv} \min\left\{2, L_v \frac{\sqrt{k\sigma_v^2 + \omega_{1-c,v}}}{k+1}\right\}, \qquad \chi(k) = \chi_1(k) + \chi_0(k). \tag{30}$$

For any $t_0 > 0$,

$$\Pr\{s(G_1) \le s(G_0)\} \le \min\left\{1, \Pr(T_\infty \le t_0) + \frac{\chi(k)^2}{t_0^2}\right\}. \tag{31}$$

For $0 < t_0 < m$, the first probability is at most $V/(V + (m - t_0)^2)$.

Proof. Using Equation (25), independence and $\mathbb{E}(\bar{y}_v - d_v) = 0$ give

$$\mathbb{E}\|q_{v,c} - d_v\|_2^2 = \frac{k\sigma_v^2 + \omega_{1-c,v}}{(k+1)^2}.$$

The direction bound in Equation (23) then implies

$$\left(\mathbb{E}\|g_\tau(q_{v,c}) - g_\tau(d_v)\|_2^2\right)^{1/2} \le \min\left\{2, L_v \frac{\sqrt{k\sigma_v^2 + \omega_{1-c,v}}}{k+1}\right\}.$$

The query G_c is independent of its own reference $q_{v,c}$. Applying the Cauchy Schwarz inequality to their inner product and factoring the expectation of their squared norms therefore bounds the error in each view. The triangle inequality in L^2 gives

$$\left(\mathbb{E}|s(G_c) - s_\infty(G_c)|^2\right)^{1/2} \le \chi_c(k).$$

Applying that triangle inequality to the difference of the query errors yields

$$\mathbb{E}|T_k - T_\infty|^2 \le \chi(k)^2.$$

If $T_\infty > t_0$ and $T_k \le 0$, the absolute margin error exceeds t_0. Markov's inequality proves Equation (31). Cantelli's inequality gives the stated bound using m and V. □

A.6 Mean Signal, Center Error, and Anomaly Coherence

The ranking bounds depend on the class margin along the population direction. Let $\mu_{cv} = \mathbb{E}_{P_c} x_v$, $\Delta_v = \mu_{1v} - \mu_{0v}$, and $a_v = \widehat{\mu}_v - \mu_{0v}$. Direct substitution gives

$$d_v = A_v(\pi\Delta_v - a_v), \qquad m = \sum_v \theta_v m_v, \qquad m_v = \frac{\Delta_v^\top P_v(\pi\Delta_v - a_v)}{\rho_v}. \tag{32}$$

The common calibration center cancels from the ideal query margin, while the calibration covariance determines its scale. To express the margin variance, concatenate the feature views and let Σ_0, Σ_1 be their full class covariance matrices, including covariance between views. Define

$$h = \mathrm{concat}_v\left(\frac{\theta_v P_v(\pi\Delta_v - a_v)}{\rho_v}\right), \qquad V = h^\top(\Sigma_0 + \Sigma_1)h. \tag{33}$$

The variance identity follows from the independent query draws. It captures class variation that remains even with an exactly known population direction.

Normal center accuracy. For the seminorm $\|z\|_{P_v} = (z^\top P_v z)^{1/2}$, the Cauchy Schwarz and reverse triangle inequalities give

$$\begin{aligned}\Delta_v^\top P_v(\pi\Delta_v - a_v) &\ge \|\Delta_v\|_{P_v}\left(\pi\|\Delta_v\|_{P_v} - \|a_v\|_{P_v}\right),\\ \kappa_v &\ge \max\{\pi\|A_v\Delta_v\|_2 - \|A_v a_v\|_2, 0\}.\end{aligned} \tag{34}$$

Thus $\|\Delta_v\|_{P_v} > 0$ and $\|a_v\|_{P_v} < \pi\|\Delta_v\|_{P_v}$ are sufficient for a positive mean margin in this view. The second inequality describes the effect of center error on reference strength. Larger reference sets reduce sampling error around d_v, while accurate normal fitting reduces the fixed shift $A_v a_v$.

Anomaly prevalence and reference size. With exact centering and $\pi > 0$, $\kappa_v = \pi\|A_v\Delta_v\|_2$. For an informative direction with $\kappa_v \ge \tau$,

$$m_v = \frac{\Delta_v^\top P_v \Delta_v}{\|A_v\Delta_v\|_2}. \tag{35}$$

The factor π cancels from this ideal standardized margin. Reference estimation still depends on the magnitude of the mixture shift. At fixed confidence, fixed geometry, and uniformly bounded reference variance, the relative sampling error scales as $1/(\pi\sqrt{k})$, and the contribution of the other query scales as $1/(\pi k)$. Holding the remaining factors fixed gives a sufficient reference size proportional to π^{-2} for a fixed relative error. Below the denominator floor, the ideal margin is $\pi\Delta_v^\top P_v\Delta_v/\tau$, so its scale also depends on prevalence.

Heterogeneous anomalies. Suppose $P_1 = \sum_{r=1}^{q} \omega_r P_{1r}$, where $\omega_r \geq 0$ and $\sum_r \omega_r = 1$. Writing $\Delta_{rv} = \mathbb{E}_{P_{1r}} x_v - \mu_{0v}$ gives $\Delta_v = \sum_r \omega_r \Delta_{rv}$. The ideal mean margin of subgroup r against a normal query is

$$m_r = \sum_v \theta_v \frac{\Delta_{rv}^\top P_v(\pi\Delta_v - a_v)}{\rho_v}. \tag{36}$$

Different anomalous substructures can reinforce the same aggregate direction. Their subgroup margins depend on their individual alignment with that direction, so an overall positive margin can coexist with a negative subgroup margin. With exact centering and balanced opposing shifts in every active view, $d_v = 0$ and the ideal score is identically zero. This loss of aggregate directional signal explains the cancellation mechanism studied in Appendix D.4.

A.7 A Controlled Comparison with Individual Distance

We use a Gaussian feature model to isolate how directional agreement and individual distance respond to irrelevant variation. The model provides an explicit comparison of the two scoring rules and the cost of estimating the reference direction.

Proposition 3 (Directional evidence under a common mean shift)**.** *Let $X_0 \sim \mathcal{N}(0, I_p)$ and $X_1 \sim \mathcal{N}(\Delta_p, I_p)$ independently, with $r_p = \|\Delta_p\|_2 > 0$, fixed $0 < \pi < 1$, exact normal centering, and identity precision and calibration covariance. The score using the population direction has*

$$\mathrm{AUROC}_{\mathrm{dir}} = \Phi(r_p/\sqrt{2}), \tag{37}$$

where Φ is the standard normal CDF. For individual squared distance $D(X) = \|X\|_2^2$, if $r_p^2 = o(\sqrt{p})$, then $\mathrm{AUROC}_{\mathrm{dist}} \to 1/2$ as $p \to \infty$. In particular, at fixed $r_p = r > 0$, directional AUROC stays above $1/2$ while distance AUROC tends to $1/2$. If $r_p \to \infty$ and $r_p^2 = o(\sqrt{p})$, directional AUROC tends to one while distance AUROC still tends to $1/2$.

Proof. The mixture displacement is $\pi\Delta_p$. For any fixed $\tau > 0$, the score $X^\top \pi\Delta_p / \max(\pi r_p, \tau)$ is a positive multiple of $X^\top \Delta_p / r_p$. The difference of the anomalous and normal projections is Gaussian with mean r_p and variance two. This proves Equation (37).

By rotational symmetry, take $\Delta_p = (r_p, 0, \ldots, 0)$. For independent standard normal vectors Z_0, Z_1, the distance margin is

$$D(X_1) - D(X_0) = \sum_{j=2}^{p} (Z_{1j}^2 - Z_{0j}^2) + (Z_{11}^2 - Z_{01}^2) + 2r_p Z_{11} + r_p^2. \tag{38}$$

The first sum divided by $2\sqrt{p-1}$ converges to a standard normal by the central limit theorem. When $r_p^2 = o(\sqrt{p})$, the remaining terms divided by the same factor converge to zero in probability. Slutsky's theorem gives $\Pr\{D(X_1) > D(X_0)\} \to 1/2$. Both score distributions are continuous, so ties have probability zero. The other conclusions follow from Equation (37). □

Estimating the direction from a finite reference. In the same Gaussian model, let $H_1, \ldots, H_k$ be independent draws from $(1-\pi)\mathcal{N}(0, I_p) + \pi\mathcal{N}(\Delta_p, I_p)$, disjoint from the queries. Put $\widehat{\delta} = k^{-1} \sum_j H_j$. The mixture covariance is $I_p + \pi(1-\pi)\Delta_p\Delta_p^\top$, so

$$\mathbb{E}\|\widehat{\delta} - \pi\Delta_p\|_2^2 = \frac{p + \pi(1-\pi)r_p^2}{k}. \tag{39}$$

For $0 < u < 1$, Markov's inequality gives the event $\|\widehat{\delta} - \pi\Delta_p\|_2 \leq u\pi r_p$ with probability at least

$$1 - \min\left\{1, \frac{p + \pi(1-\pi)r_p^2}{ku^2\pi^2 r_p^2}\right\}.$$

On this event, $\widehat{\delta} \neq 0$, and the conditional AUROC for a common reference satisfies

$$\mathrm{AUROC}(\widehat{\delta}) = \Phi\left(\frac{\Delta_p^\top \widehat{\delta}}{\sqrt{2}\|\widehat{\delta}\|_2}\right) \geq \Phi\left(\frac{r_p}{\sqrt{2}} \frac{1-u}{1+u}\right). \tag{40}$$

The equality follows from the conditional Gaussian projection distributions. For the inequality, use

$$\Delta_p^\top \widehat{\delta} \geq \pi r_p^2(1-u), \qquad \|\widehat{\delta}\|_2 \leq \pi r_p(1+u).$$

A positive common denominator preserves the conditional ranking. The probability bound shows how increasing reference size can compensate for higher dimension or lower prevalence at a given signal strength.

A.8 Effect of View Fusion on Ranking

Size information can improve detection when it adds class separation beyond that already present in structural scores. Hold the fitted scoring rules fixed and use a common external reference in each view. Let T_s, T_a be the anomalous score minus the normal score in the two views, with finite means m_s, m_a, variances V_s, V_a, and covariance C_{sa}. Since the factor $1/(1+\lambda)$ preserves ranking, the relevant margin moments are

$$m(\lambda) = m_s + \lambda m_a, \qquad V(\lambda) = V_s + 2\lambda C_{sa} + \lambda^2 V_a. \tag{41}$$

For $m(\lambda) > 0$, Cantelli's inequality gives

$$\Pr\{T_s + \lambda T_a \leq 0\} \leq \frac{V(\lambda)}{V(\lambda) + m(\lambda)^2}. \tag{42}$$

Assume $m_s > 0$, $V_s > 0$, $\lambda \geq 0$, $m(\lambda) > 0$, and $V(\lambda) > 0$. Fusion gives a strictly smaller upper bound than the structural view exactly when

$$2\lambda m_s(m_a V_s - m_s C_{sa}) + \lambda^2(m_a^2 V_s - m_s^2 V_a) > 0. \tag{43}$$

To verify this condition, compare $m(\lambda)^2/V(\lambda)$ with m_s^2/V_s and expand $V_s(m_s + \lambda m_a)^2 - m_s^2(V_s + 2\lambda C_{sa} + \lambda^2 V_a)$.

If the two margins are jointly Gaussian, the ranking probability is

$$\Pr(T_s + \lambda T_a > 0) = \Phi\left(\frac{m_s + \lambda m_a}{\sqrt{V_s + 2\lambda C_{sa} + \lambda^2 V_a}}\right). \tag{44}$$

In this case, Equation (43) is also the exact condition for improving the ranking probability. Under the general moment assumptions, it compares the error bounds in Equation (42).

The effect of a small auxiliary weight follows from

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{m_s + \lambda m_a}{\sqrt{V_s + 2\lambda C_{sa} + \lambda^2 V_a}}\bigg|_{\lambda=0} = \frac{m_a V_s - m_s C_{sa}}{V_s^{3/2}}. \tag{45}$$

Thus $m_a V_s > m_s C_{sa}$ makes sufficiently small positive weights improve the margin ratio and its error bound. Under the Gaussian model, this also improves the ranking probability. The quadratic term in Equation (43) explains why larger weights can reduce the benefit.

This complementary information has a direct interpretation through the residual size margin

$$T_\perp = T_a - \frac{C_{sa}}{V_s} T_s, \qquad \mathbb{E}T_\perp = m_a - \frac{C_{sa}}{V_s} m_s, \qquad \mathrm{Var}(T_\perp) = V_a - \frac{C_{sa}^2}{V_s}. \tag{46}$$

The residual is uncorrelated with T_s, and joint Gaussianity makes them independent. The local improvement condition is $\mathbb{E}T_\perp > 0$, meaning that the size view retains positive class separation after accounting for its linear relation to the structural margin.

PopGLAD selects $\lambda = \sqrt{\alpha}$ from normal structural moments, giving $\theta_a = \lambda/(1+\lambda) \leq 1/2$. This rule regulates the auxiliary contribution using statistics available during fitting. Its benefit depends on the class margins and their covariance, as described by Equation (43). The component experiments in Appendix D.1 evaluate these effects for the chosen rule.

The weight also gives a direct bound on score changes. If $\|b_v(G)\|_2 \leq R_v$, then $|s_v(G)| \leq R_v$ for any reference, and

$$\left|\frac{s_s(G) + \lambda s_a(G)}{1+\lambda} - s_s(G)\right| \leq \frac{\lambda}{1+\lambda}(R_s + R_a). \tag{47}$$

The corresponding change in a query pair's margin is at most twice this quantity. This bound applies to both common references and the references used by PopGLAD, with the fitted view scores held fixed.

B Algorithm and Computational Complexity

B.1 Feature Construction

Categorical node attributes, edge attributes, and refinement tuples use a deterministic 64-bit hash. The refinement tuple contains the current node label, incoming degree, and a sorted multiset of edge-label and neighbor-label pairs. Each depth occupies a separate histogram block. The final coordinate is the hash modulo the block width. Consequently, graph identity and ordering have no role in the feature map. Finite hashing can merge distinct patterns, and ordinary neighborhood refinement can assign the same representation to different graphs.

The size input is

$$u(G) = \left(\log(1 + |V_G|), \log(1 + |E_G|), |V_G|^{-1} \sum_{r \in V_G} \mathbf{1}\{\deg(r) > 0\} \right). \tag{48}$$

For each coordinate, apply the inverse hyperbolic sine and subtract its fitting mean. Divide by its fitting sample standard deviation with a floor of 0.05. The three dimensional covariance uses the Ledoit–Wolf estimator, with its Moore–Penrose inverse when singular.

B.2 Fitting and Scoring Algorithm

Algorithm 1 Normal fitting and population scoring in PopGLAD

Require: Disjoint normal fitting and calibration sets $\mathcal{F}, \mathcal{C}$, with $|\mathcal{F}| \geq 4$ and $|\mathcal{C}| \geq 2$
Count attributed neighborhoods at depths $0, \ldots, L$
Fit structural frequency weights and the size transformation on $\mathcal{F}$
Compute $\widehat{\mu}_s$, coordinate variances, and fourth moments on $\mathcal{F}$
Compute α, η_s, P_s using Equations (3) and (4)
Fit $\widehat{\mu}_a, P_a, \eta_a$ on the size representations of $\mathcal{F}$
Select K calibration graphs and compute c_v and the centered calibration matrices $Z_{\mathcal{C},v}$ for both views
Set $\theta_s = 1/(1 + \sqrt{\alpha})$ and $\theta_a = 1 - \theta_s$
Freeze all fitting and calibration quantities
Require: A fixed unlabeled collection $\mathcal{U}$ with $n \geq 2$
Transform each graph independently into $x_s(G_i)$ and $x_a(G_i)$
for each view v **do**
    Compute $\bar{z}_v = n^{-1} \sum_i x_v(G_i) - \widehat{\mu}_v$
    Compute the shared products $P_v \bar{z}_v$ and $Z_{\mathcal{C},v} P_v \bar{z}_v$
    for each computation batch **do**
        Evaluate the query numerators by Equation (10)
        Evaluate directional variances by Equation (50)
        Divide each numerator by its standard deviation, with floor τ
    end for
end for
return $s_i = \theta_s s_{s,i} + \theta_a s_{a,i}$ for each graph identity

The same calibration identities are used in both views. The implementation uses float64 for moments and scores. Variances and quadratic forms are clamped at zero when floating point cancellation produces a small negative value. These numerical corrections leave the exact arithmetic definitions unchanged.

B.3 Efficient Scoring with Query Exclusion

For one view, suppress the view subscript and put $z_i = x_i - \widehat{\mu}$ and $\bar{z} = n^{-1} \sum_i z_i$. Let $Z_{\mathcal{C}} \in \mathbb{R}^{K \times d}$ denote the centered calibration matrix whose rows are $x(G) - c$ for the selected K calibration graphs, so that $C = Z_{\mathcal{C}}^{\top} Z_{\mathcal{C}} / (K - 1)$. For each query,

$$\delta_{-i} = \frac{n\bar{z} - z_i}{n - 1}, \qquad Z_{\mathcal{C}} P \delta_{-i} = \frac{n Z_{\mathcal{C}} P \bar{z} - Z_{\mathcal{C}} P z_i}{n - 1}. \tag{49}$$

The squared denominator before applying the floor is

$$\begin{aligned}\delta_{-i}^{\top} P\Omega P\delta_{-i} = \frac{\|nZ_{\mathcal{C}}P\bar{z} - Z_{\mathcal{C}}Pz_i\|_2^2}{(K-1)(n-1)^2} \\ + \frac{\eta}{(n-1)^2}\left(n^2\bar{z}^{\top}P^2\bar{z} - 2nz_i^{\top}P^2\bar{z} + z_i^{\top}P^2z_i\right).\end{aligned} \tag{50}$$

Together with Equation (10), this gives the exact score using shared collection summaries and products between a query and the calibration graphs.

For diagonal $P = \mathrm{diag}(p)$, dense centering is avoided using

$$z_i^{\top}Pz_i = (x_i \odot x_i)^{\top}p - 2x_i^{\top}P\widehat{\mu} + \widehat{\mu}^{\top}P\widehat{\mu}. \tag{51}$$

Replacing p with $p \odot p$ yields $z_i^{\top}P^2z_i$. The calibration cross product also expands into sparse products and fixed centering terms. Neither a dense matrix of centered structural features nor an $n \times n$ graph similarity matrix is required.

Let S count the stored entries of all supplied structural features, d their dimension, and K the calibration budget. Moment fitting costs $O(S+d)$. A conservative bound for the calibration products and all scores is $O(KS+Kd+nK)$. Storing features, fitted vectors, and scores requires $O(S+d+n)$ memory. Computation batches of size B add $O(BK)$ workspace. With the fixed calibration budget, arithmetic and storage grow linearly with the sparse feature input and dimension.

B.4 Feature Extraction Complexity

Let N and E denote the total number of nodes and directed edges in a graph collection, with fixed attribute widths. Incoming edges are grouped by their target node through sorting. Neighborhood refinement sorts the edge-label and neighbor-label pairs at each node. The implementation uses insertion sort, costing at most $O(\sum_r \deg(r)^2)$ per depth. Histogram construction sorts labels within each graph and depth. A conservative extraction bound is

$$O\left(N + E\log(E+1) + L\left(N + E + \sum_r \deg(r)^2\right) + (L+1)\sum_G |V_G|\log(|V_G|+1)\right). \tag{52}$$

The sparse feature count satisfies $S \le (L+1)N$. Token storage costs $O((L+1)N)$ in addition to the adjacency and histogram arrays. For molecular graphs with bounded degree, the degree term is linear in N.

C Experimental Details

Datasets. Table 5 summarizes the detailed information of the eight graph benchmarks used in our experiment. Note that for each dataset, we treat the majority class as normal and the minority class as anomalous. For the data split, we use 80% of the normal graphs as the training pool and the remaining normal graphs for testing. Specifically, for ER_MD and AIDS, we select fixed subsets of anomalous graphs to form balanced test sets. For the other datasets, all anomalous graphs are included in the test set. All methods in the experiment share the same test partitions to ensure a fair comparison. The anomaly fractions in Table 5 refer to the complete datasets before preprocessing and splitting.

Implementation details. In PopGLAD, 80% of the graphs within the training pool are used for fitting, and 20% are reserved for calibration. The test set remains fixed across runs. The fitting set determines the feature transformations, normal centers, precision estimates, and fusion weights. PopGLAD uses refinement depth $L = 4$, hash width $M = 2^{20}$ per depth, and up to 128 calibration graphs. Note that we use the entire calibration pool when it contains fewer than 128 graphs. The same calibration graphs are used for both views to estimate the calibration mean and covariance. The structural base ridge is 32 times the mean fitted coordinate variance, with shrinkage toward the mean variance over active coordinates. The size ridge is 0.01 times the mean fitting variance. The moment rule determines the shrinkage and fusion weights. These settings are shared across datasets. We use categorical node and edge attributes, assigning a constant category to untyped edges. During inference, each query uses all other test graphs as its population reference.

Table 5: Dataset statistics for the datasets used in the experiment.

Dataset	# Graphs	#Mean nodes	# Mean edges	# Graph classes	#Node classes	Anomaly rate (%)
MUTAG	188	17.93	19.79	2	7	33.51
AIDS	2,000	15.69	16.20	2	38	20.00
DD	1,178	284.32	715.66	2	82	41.34
ER_MD	446	21.33	234.85	2	10	40.58
SW-620	40,532	26.06	28.09	2	65	5.95
MOLT-4	39,765	26.10	28.14	2	64	7.90
PC-3	27,509	26.36	28.49	2	45	5.70
MCF-7	27,770	26.40	28.53	2	46	8.26

Baseline methods. We use the UB-GOLD implementations (Wang et al., 2025) to execute baselines including GLocalKD, OCGTL, OCGIN, SIGNET, CVTGAD, GLADC, GOOD-D, and GraphDE. To implement baselines such as MUSE, GUL, and GLADPro, we use the officially released codes from the authors. All baselines use the same normal training pool and test partitions as PopGLAD. Their hyperparameters follow the settings reported in their respective papers.

Ablation variants. The variants evaluated in Section 5.3 and Appendix D.1 share PopGLAD's graph representations. With $z_i = x_s(G_i) - \widehat{\mu}_s$ and $P_{\text{fix}} = \text{diag}(v_j + \gamma_0)^{-1}$, individual distance is $z_i^\top P_s z_i$, the mean witness is $z_i^\top \delta_{s,-i}$, and the precision witness is $z_i^\top P_{\text{fix}} \delta_{s,-i}$. The fixed covariance variant sets $P_s = P_{\text{fix}}$ and $\eta_s = \gamma_0$, while retaining the fitted fusion weights. The structural view and size view variants use their respective calibrated scores. Equal fusion averages the two calibrated view scores.

Additional experimental settings. We use SW-620, MOLT-4, PC-3, and MCF-7 for the additional experiments on graph datasets. To study the reference population, we keep the query graphs and fitted model fixed and sample reference graphs from a pool disjoint from the queries. Sampling within each class allows us to vary the reference size or anomaly fraction (Appendix D.2). For calibration sensitivity, we vary the number of normal calibration graphs while keeping the fitting statistics and query graphs fixed (Appendix D.5). To study structural sensitivity, we compare each original graph with a version obtained by edge swaps that preserve graph size and node degrees. Both graphs are scored using the same external reference and normal statistics. A separate synthetic experiment examines cancellation in the reference mean by comparing a common anomaly mean shift with two anomaly groups having equal weights and opposite mean shifts (Appendix D.4). To assess numerical consistency, we change the query order and computation batch size while keeping the graph collection and fitted model fixed (Appendix D.6).

Runtime evaluation. We use stratified five-fold partitions for runtime evaluation on an Xeon Platinum 8480CL CPU and H100 80GB GPUs. Timings cover data loading, graph preparation, fitting, and scoring. Note that training parameters, such as training epochs, may differ across methods due to their training protocols and convergence rates. Here we provide the detailed settings for some baselines. For example, GLocalKD uses 150 epochs. OCGTL, OCGIN, and SIGNET retain their stopping rules with caps of 500, 500, and 1000 epochs. GUL and GLADPro use ten epochs. MUSE uses ten encoder epochs and 100 classifier epochs. Other baselines not specifically mentioned follow their default settings.

Table 6: Anomaly detection performance using different covariance estimation and score fusion (mean $\pm$ standard deviation (std), %). Bold indicates the best results.

Method	SW-620		MOLT-4		PC-3		MCF-7	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Fixed covariance	87.36 ± 0.08	76.53 ± 0.06	83.68 ± 0.11	76.28 ± 0.13	87.32 ± 0.09	76.18 ± 0.12	85.20 ± 0.17	78.43 ± 0.19
Structural view	86.00 ± 0.08	75.19 ± 0.06	83.31 ± 0.12	76.05 ± 0.12	86.41 ± 0.07	75.45 ± 0.12	85.15 ± 0.07	78.23 ± 0.11
Size view	66.97 ± 0.12	40.53 ± 0.17	64.75 ± 0.06	44.03 ± 0.10	68.52 ± 0.03	41.91 ± 0.07	65.91 ± 0.04	48.86 ± 0.05
Equal fusion	87.04 ± 1.20	76.46 ± 1.72	82.18 ± 1.56	74.43 ± 2.04	87.49 ± 0.83	76.40 ± 1.33	84.81 ± 0.96	76.26 ± 1.24
PopGLAD	$\mathbf{89.24 \pm 0.10}$	$\mathbf{79.35 \pm 0.09}$	$\mathbf{85.44 \pm 0.17}$	$\mathbf{78.19 \pm 0.16}$	$\mathbf{88.87 \pm 0.11}$	$\mathbf{78.77 \pm 0.18}$	$\mathbf{86.76 \pm 0.08}$	$\mathbf{80.79 \pm 0.12}$

D Additional Experimental Results

D.1 Covariance Estimation and View Fusion

To examine the contributions of view fusion and covariance estimation, we compare PopGLAD with the component baselines in Table 6: (1) the ***Fixed covariance*** baseline sets $P_s = \mathrm{diag}(v_j + \gamma_0)^{-1}$ and $\eta_s = \gamma_0$, while retaining the fitted fusion weights, (2) the ***Structural view*** and ***Size view*** baselines use their respective calibrated view scores, and (3) the ***Equal fusion*** assigns a weight of one-half to each view.

On MOLT-4, the structural view achieves 83.31% AUROC and the size view achieves 64.75%. Combining the views with the proposed weighting improves AUROC over the structural view by 2.13 percentage points. Thus, size information complements structural scores even though it gives weaker detection on its own. The fusion weight determines whether this complementary information improves detection. Equal fusion reduces AUROC on MOLT-4 by 1.13 percentage points relative to the structural view. The proposed weighting gives the structural view greater influence and achieves 85.44% AUROC. The two comparisons show that a view can help at a modest weight while reducing performance when given equal influence.

With the fusion weight fixed, covariance estimation improves both metrics over the fixed covariance control on all four graph datasets. On SW-620, the AUROC gain is 1.88 percentage points, even though the mean shrinkage coefficient is only 0.0244 (Table 7). Here covariance estimation affects both structural precision weighting and directional calibration. Its influence, therefore, includes how structural deviations are weighted and how their scores are balanced with the size view. The gain supports this joint role of covariance estimation in the final score.

Table 7: Shrinkage coefficient α (mean $\pm$ std) and auxiliary weight $\sqrt{\alpha}$ (mean) estimated from normal graphs over ten runs.

Quantity	SW-620	MOLT-4	PC-3	MCF-7
α	0.0244 ± 0.0001	0.0251 ± 0.0001	0.0329 ± 0.0001	0.0333 ± 0.0001
$\sqrt{\alpha}$	0.1561	0.1583	0.1813	0.1825

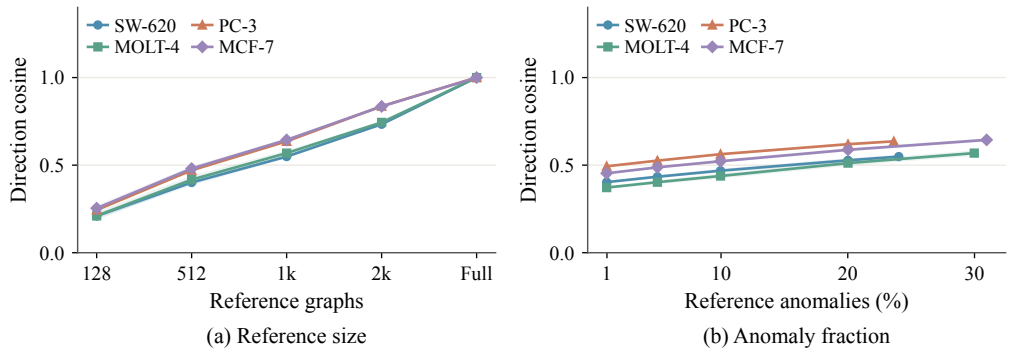

Figure 5: Reference direction stability. Cosine similarity to the precision-weighted structural direction of the complete reference pool under (a) varying reference size at the pool's anomaly fraction and (b) varying anomaly fraction with 1024 reference graphs. Lines and bands show mean $\pm$ population standard deviation over ten resamples. "Full" denotes the complete pool. The final point in (b) uses each pool's anomaly fraction.

D.2 Reference Direction Stability

To examine the stability of the direction used for scoring, we measure cosine similarity between sampled and complete reference directions under the protocol of Section 5.4. The compared directions are structural mean displacements multiplied by the fitted precision P_s. The fitted statistics and query set remain fixed, and queries are disjoint from the reference pool. In Figure 5(a), increasing reference

size from 128 to 2048 raises cosine similarity from 0.21–0.25 to 0.73–0.84 across the four datasets. Larger samples therefore yield directions closer to that of the complete pool. Similarity at "Full" is one by construction.

With 1024 reference graphs, increasing the anomaly fraction also raises direction similarity on all four datasets (Figure 5(b)). This diagnostic measures agreement with the complete pool, whose direction is itself estimated from a finite collection. It complements the detection results in Figure 2 and is consistent with the role of reference uncertainty in Theorem 1.

D.3 Score Analysis on PC-3 and MCF-7

We extend the score analysis in Figure 3 to PC-3 and MCF-7. In Figure 6, normal scores concentrate near zero, while anomaly scores have higher medians and longer positive tails. For PC-3, the normal and anomaly medians are -0.12 and 2.94, respectively. For MCF-7, they are -0.05 and 2.55. These additional datasets exhibit separation in typical scores as well as stronger responses from a subset of anomalies.

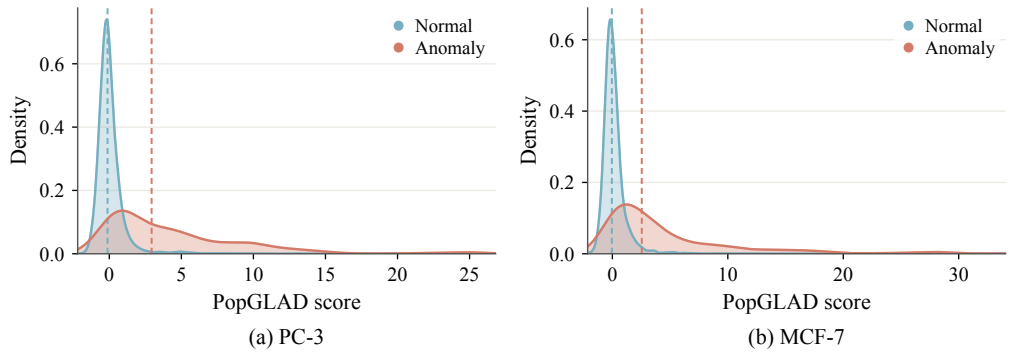

Figure 6: Class conditional PopGLAD score densities on (a) PC-3 and (b) MCF-7, complementing SW-620 and MOLT-4 in Figure 3(a,b). Dashed lines mark class medians. Each density is normalized separately and includes the full observed score range from one run per dataset (seed 55).

Figure 7 compares individual distances and PopGLAD scores for the same queries. In the visualized runs, 93.49% of PC-3 anomalies and 93.72% of MCF-7 anomalies fall below the 95th percentile of normal individual distances. A threshold at this percentile would miss most anomalies in both datasets, although many receive high population scores. Together with their individual distance controls in Table 3, these additional results support the value of directional agreement when normal and anomalous graphs occupy overlapping distance ranges.

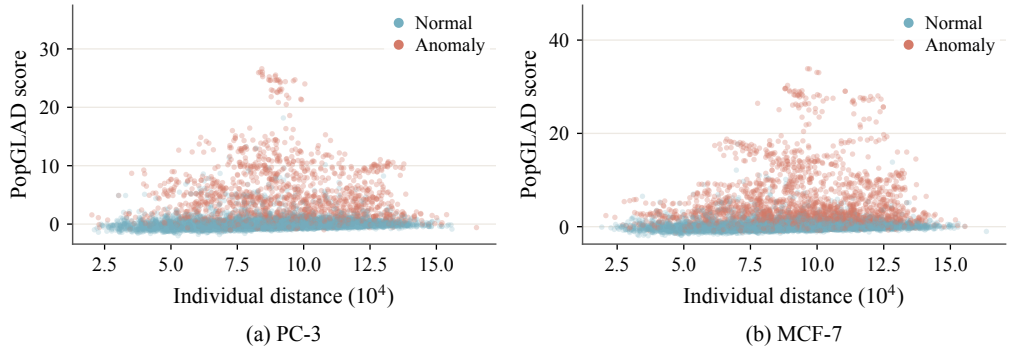

Figure 7: Individual distance and PopGLAD score for every query on (a) PC-3 and (b) MCF-7, using the same runs as Figure 6. These panels complement SW-620 and MOLT-4 in Figure 3(c,d). All query points are retained.

D.4 Cancellation between Opposing Anomaly Shifts

Beyond the graph-edit analysis in Section 5.5, we examine whether individual anomaly signals remain useful after aggregation. We compare shared and balanced opposing shifts in a synthetic feature space experiment. The shared condition gives anomalies a common mean shift. The opposing condition uses two equally weighted anomaly groups with opposite mean shifts.

Table 8 reports 100% AUROC for shared shifts and 50.11% for opposing shifts. Shared deviations reinforce the reference mean, while opposing deviations cancel, even though individual anomalies remain far from normal inputs. The result distinguishes information in individual deviations from information preserved by their mean. Directional coherence determines whether deviations reinforce one another in the reference, while reference size controls the sampling error in estimating that mean. This distinction is consistent with Equation (9).

Table 8: Detection performance under shared and opposing anomaly shifts in a synthetic feature space. Mean ± std (%) over thirty paired repetitions.

Population shift	AUROC	AUPRC
Shared	100.00 ± 0.00	100.00 ± 0.00
Opposing	50.11 ± 0.74	69.87 ± 3.62

D.5 Sensitivity to Calibration Size

To evaluate sensitivity to calibration size, we vary the number of normal calibration graphs within the same fitted model in Figure 8. The default budget of 128 graphs differs from the complete calibration pool by at most 0.042% in AUROC and 0.033% in AUPRC across all evaluated graph datasets. For a single view with fixed precision and a common reference direction, calibration applies a positive scale and a common offset, preserving the ranking of graphs. In PopGLAD, the directions vary across queries, and calibration also determines the relative scales of the two views. The small metric differences show that these effects have a limited influence across the tested budgets, which supports the default calibration size for efficient scoring.

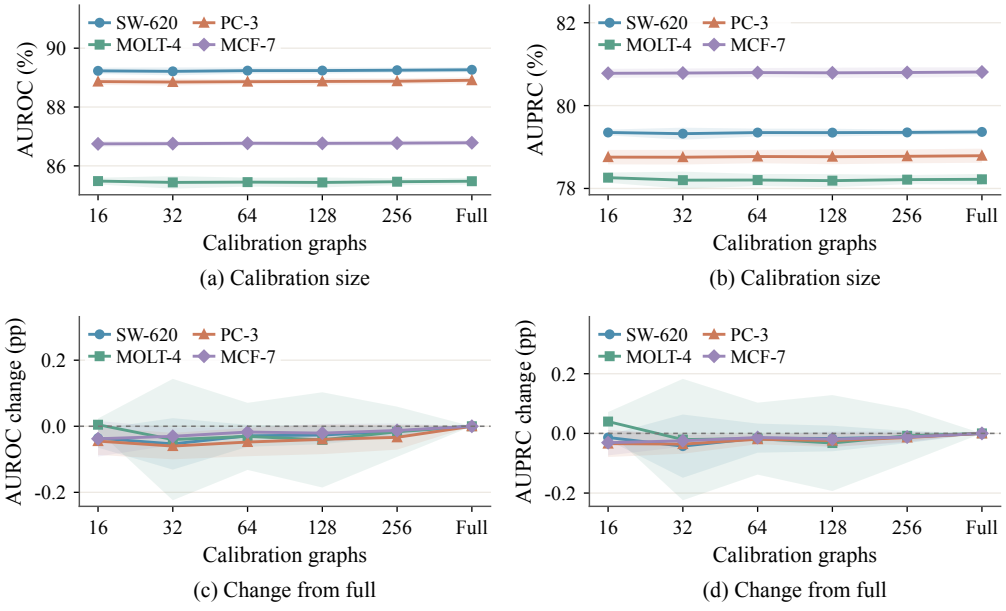

Figure 8: Effect of calibration set size on detection performance. (a,b) AUROC and AUPRC. (c,d) Paired differences from full calibration in percentage points. Lines and bands show mean ± std over ten normal partitions. Full denotes the complete calibration pool.

D.6 Invariance to Query Order and Batching

To verify numerical consistency, we vary query order and computation batch size within a fixed graph collection. Table 9 reports identical AUROC and AUPRC in every tested setting, with the largest absolute score difference of 1.9×10^{-12} on PC-3. These results agree with the collection equivariance in Proposition 1. Reordering queries or changing computation batches preserves the reference available to each graph. Batching can therefore be chosen to fit available memory while retaining the same population context and detection performance.

Table 9: Maximum absolute changes in scores and detection metrics under different query orders and computation batch sizes. Metric changes are in percentage points.

Dataset	Maximum score change	Maximum metric change
SW-620	7.0×10^{-13}	0.00
MOLT-4	5.2×10^{-13}	0.00
PC-3	1.9×10^{-12}	0.00
MCF-7	3.5×10^{-13}	0.00

D.7 Screening with Limited Inspection Budgets

To evaluate detection under limited inspection budgets, we measure anomaly recall among the graphs with the highest scores in Figure 9. At a 10% inspection budget on PC-3, PopGLAD achieves 38.87% recall, exceeding the strongest comparison method, GOOD-D, by 17.21 percentage points. This is approximately 3.89 times the recall expected under random inspection, indicating a much higher concentration of anomalies in the selected graphs. At a 20% budget, recall reaches 67.45%. The additional recovery shows that useful anomaly evidence extends beyond the first 10% of graphs, allowing greater coverage as inspection resources increase.

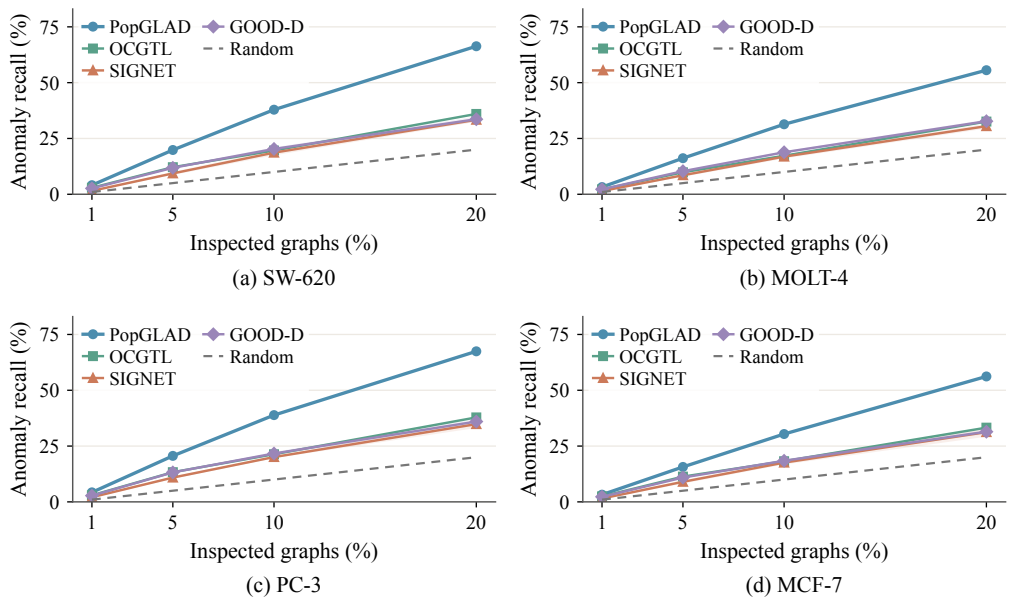

Figure 9: Anomaly recall at different inspection budgets. Lines and bands show mean $\pm$ std over ten runs. The dashed line shows expected recall under random ranking.