SCOPE: SELF-CONSTRUCTED POSITIVE EVIDENCE FOR IN-CONTEXT TABULAR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular anomaly detection is challenging because anomalies are rare, diverse, and typically unlabeled. Existing methods largely model normality, while recent tabular foundation models either rely on anomaly-specific pretraining with synthetic anomalies or use predictive discrepancy directly as the anomaly score. What remains largely unexplored is whether a general-purpose tabular foundation model can construct the supervision needed to use its discriminative in-context capability at inference time. We introduce **SCOPE** (**S**elf-**CO**nstructed **P**ositive **E**vidence), which turns a frozen general-purpose tabular foundation model into an anomaly detector without anomaly-specific training. SCOPE first uses conditional feature prediction to locate candidate positive evidence, then uses discriminative in-context prediction to iteratively refine this evidence. Uncertain target rows receive graded intermediate targets rather than being treated as normal, and multiple evidence contexts are sampled from the refined ranking to reduce dependence on a single hard positive set. All context rows are actual observations rather than synthesized anomalies, and the frozen predictor is never updated. Across 1,502 benchmark entries from ADBench, OddBench, and OvRBench, SCOPE ranks first under all five aggregate criteria for both AUROC and AUPRC on every benchmark.

1 INTRODUCTION

Tabular anomaly detection requires distinguishing anomalous from normal instances without labeled anomalies. We consider the one-class setting, where normal training observations are available but anomalous examples are not. Classical methods define normality through quantities such as distance, density, or partition depth: kNN uses nearest-neighbor distance (Ramaswamy et al., 2000), LOF local density ratios (Breunig et al., 2000), Isolation Forest partition depth (Liu et al., 2008), and OCSVM a boundary enclosing normal data (Schölkopf et al., 1999). Deep methods instead learn such structures from data (Ruff et al., 2018; Bergman & Hoshen, 2020; Shenkar & Wolf, 2022; Thimonier et al., 2024; Livernoche et al., 2024). Yet without supervision indicating what constitutes an anomaly, learning the appropriate notion of normality remains difficult.Indeed, large-scale evaluations show that sophisticated learning-based methods do not consistently dominate simple classical baselines such as nearest-neighbor methods (Han et al., 2022).

Recent tabular foundation models offer a new opportunity (Hollmann et al., 2023; Qu et al., 2026; Grinsztajn et al., 2026). Some approaches pretrain anomaly-specific priors using synthetic anomaly-detection tasks (Shen et al., 2025; Ding et al., 2026b), but their generalization can depend on how well deployment anomalies align with the synthetic priors used during pretraining. Other work converts predictive discrepancies from general-purpose tabular foundation models directly into anomaly scores. For example, TabPFN-OD adapts TabPFN (Hollmann et al., 2023) through feature-wise self-prediction (Ding et al., 2026a). We ask a different question: *can we exploit the existing discriminative capability of a general-purpose tabular foundation model that was never trained specifically for anomaly detection?* Given labeled in-context examples, such models can solve new discrimination tasks without parameter updates. Although anomalous examples are unavailable in the one-class setting, **the evidence needed to construct such a context may already be present in the unlabeled target data**.

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.

We introduce **SCOPE** (**S**elf-**CO**nstructed **P**ositive **E**vidence). Rather than using conditional prediction from a frozen tabular foundation model as the final anomaly score, SCOPE uses it only to locate target rows that may provide positive evidence. This evidence is then re-evaluated through the model's discriminative in-context capability, and the resulting scores refine which target rows provide stronger evidence. Crucially, SCOPE does not force the target pool into a hard normal-versus-anomalous partition. Trusted normal references receive target 0, uncertain target rows receive graded intermediate targets according to their refined evidence, and rows with stronger evidence are more likely to be sampled as positive context with target 1. All context rows are actual observations rather than synthesized anomalies.

Because the positive evidence is uncertain, SCOPE avoids committing to a single hard set of provisional anomalies. After refinement, it converts the evidence ranking into a sampling distribution and draws multiple provisional-positive sets, favoring rows with stronger evidence. Their scores are then aggregated. Cross-fitting prevents a target row from directly reinforcing its own provisional evidence, while held-out normal references align scores obtained from different contexts. The tabular foundation model remains frozen throughout: constructed targets condition its in-context predictions but never update its parameters. Thus, SCOPE requires neither anomaly-specific pretraining nor task-specific fine-tuning.

Across three large-scale tabular anomaly-detection benchmarks comprising $1{,}502$ benchmark entries, SCOPE consistently outperforms classical, deep, and foundation-model-based baselines, ranking first under all five reported aggregate criteria for both AUROC and AUPRC on each benchmark. Our analyses further show that the main improvement arises not from using predictive signals directly as anomaly scores, but from converting them into discriminative in-context supervision; cross-fitted refinement then further corrects the resulting evidence rather than simply preserving the initial ranking. These findings suggest a different role for predictive modeling in anomaly detection: predictive signals can serve as evidence from which a pretrained discriminative model constructs its own anomaly-detection context.

Our main contributions are:

- **From predictive signals to positive evidence.** We show that signals derived from conditional feature prediction are more useful for locating candidate positive evidence than for directly ranking anomalies. This reframes predictive modeling from a final scoring mechanism into a means of constructing supervision for downstream discrimination.
- **Self-constructed discrimination with a frozen tabular foundation model.** Building on this view, we introduce SCOPE, which constructs a discriminative in-context task directly from normal references and an unlabeled target table. The framework requires neither anomaly-specific pretraining nor parameter updates, while cross-fitted and uncertainty-aware evidence construction prevents uncertain observations from being prematurely committed as either normal or anomalous.
- **Large-scale empirical validation and mechanistic analysis.** We evaluate SCOPE on $1{,}502$ benchmark entries across ADBench, OddBench, and OvRBench, where it achieves the best performance among all compared methods under all five reported aggregate criteria for both AUROC and AUPRC on every benchmark. Our analyses further isolate the contributions of discriminative in-context prediction, cross-fitted refinement, and evidence re-ranking, showing that the gains do not arise from the initial predictive signal alone.

2 Related Work

2.1 Tabular Anomaly Detection

Classical tabular anomaly detectors characterize abnormality through geometric or statistical properties of the observed data. Representative approaches use nearest-neighbor distance (Ramaswamy et al., 2000), local density ratios (Breunig et al., 2000), random-partition depth (Liu et al., 2008), or one-class decision boundaries (Schölkopf et al., 1999). Deep methods instead learn representations or scoring functions for normality, including compact one-class representations (Ruff et al., 2018), self-supervised objectives (Bergman & Hoshen, 2020; Shenkar & Wolf, 2022), inter-sample modeling (Thimonier et al., 2024), and diffusion-based approaches (Livernoche et al., 2024). arge-

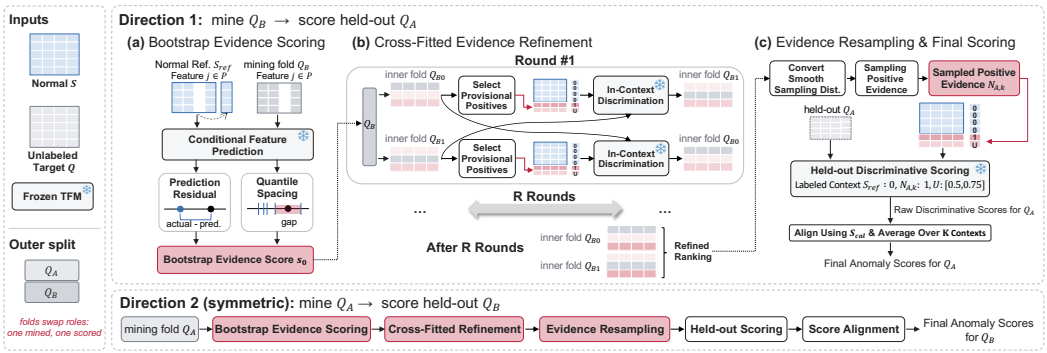


Figure 1: Overview of SCOPE.

scale evaluations nevertheless show that sophisticated learning-based methods do not consistently outperform strong classical baselines, including nearest-neighbor methods (Han et al., 2022). A complementary line introduces surrogate positive evidence rather than relying solely on normality modeling. Outlier Exposure uses samples from auxiliary datasets (Hendrycks et al., 2019), while CutPaste constructs artificial anomalies through domain-specific transformations (Li et al., 2021). For tabular data, SO-GAAL generates candidate anomalies adversarially (Liu et al., 2019), DROCC searches for informative perturbed samples around observed data (Goyal et al., 2020), and AutoAnoEval uses LLM-generated pseudo-anomalies as a pseudo-evaluation set for unsupervised model selection (Yoon et al., 2026). These approaches demonstrate the value of introducing positive evidence, but that evidence is supplied externally or constructed through a predefined generation or perturbation mechanism. SCOPE instead selects provisional positive evidence directly from the target table and uses it only to condition a frozen predictor rather than to update its parameters. The selected labels are imperfect and are therefore treated as evidence to be re-evaluated, rather than as ground-truth anomalies.

2.2 Tabular Foundation Models for Anomaly Detection

Recent work has begun to exploit tabular foundation models for anomaly detection. FoMo-0D (Shen et al., 2025) pretrains a foundation model on synthetic anomaly-detection tasks, OutFormer (Ding et al., 2026b) extends this paradigm with richer synthetic priors and curriculum training, and ICLAD (Wei & Armanfard, 2026) meta-learns across synthetic anomaly-detection tasks to support multiple supervision regimes. These methods show that anomaly-specific supervision can be incorporated during pretraining and transferred to unseen datasets. Because this supervision is synthesized before the target dataset is observed, however, the anomaly concepts available to the pretrained predictor are shaped by the chosen synthetic priors. A different direction keeps a general-purpose tabular foundation model frozen and derives anomaly scores from its predictive behavior. TabPFN-OD adapts TabPFN (Hollmann et al., 2023) through feature-wise self-prediction (Ding et al., 2026a), while uLEAD-TabPFN (Lu et al., 2026) scores violations of conditional feature dependencies with uncertainty-aware aggregation. In both cases, predictive discrepancy is used to construct the anomaly score. SCOPE also starts from a frozen general-purpose model, but uses this discrepancy only to bootstrap provisional positive evidence selected from the target table. The resulting labeled context then invokes the model's existing discriminative in-context capability, without anomaly-specific pretraining or parameter updates.

3 Method

3.1 Setup and Overview

Let $\mathcal{S}_{\text{all}}$ denote the available normal training observations and

$$\mathcal{Q} = \{x_i\}_{i=1}^{m} \subset \mathbb{R}^d$$

the unlabeled target observations to be scored. We sample a fixed normal pool from $\mathcal{S}_{\text{all}}$ and divide it into a reference set $\mathcal{S}_{\text{ref}}$ and a smaller calibration set $\mathcal{S}_{\text{cal}}$. The former is used to construct prediction

and discrimination contexts, whereas the latter is held out for final score alignment. Figure 1 summarizes the complete SCOPE pipeline, including the outer split, bootstrap evidence construction, cross-fitted refinement, evidence resampling, and final scoring.

To ensure that a target observation does not contribute to the evidence used to score itself, we split $\mathcal{Q}$ into two disjoint outer folds, $\mathcal{Q}_A$ and $\mathcal{Q}_B$. When scoring fold $\mathcal{Q}_b$, positive evidence is constructed exclusively from the opposite fold, denoted by $\mathcal{M}_b$. Thus, $\mathcal{M}_A = \mathcal{Q}_B$ and $\mathcal{M}_B = \mathcal{Q}_A$.

SCOPE then proceeds in three stages. First, it uses *conditional feature prediction* to obtain an initial evidence ranking within $\mathcal{M}_b$. A small subset of feature columns is selected, and each selected feature is masked in turn and predicted from the remaining features using normal reference examples. Signals from these predictions identify target observations that may provide positive evidence.

Second, this evidence is converted into an *in-context regression* task. Known normal observations receive target 0, selected positive-evidence observations receive target 1, and other target observations receive intermediate values rather than being treated as normal. Cross-fitted discrimination then re-evaluates and refines the initial evidence ranking.

Finally, SCOPE samples multiple positive-evidence sets from the refined ranking, scores the held-out query fold under each resulting context, aligns the scores using $\mathcal{S}_{\text{cal}}$, and averages them to obtain the final anomaly score.

Conditional feature prediction and in-context discrimination therefore play distinct roles: the former locates where positive evidence may lie, while the latter determines how that evidence is translated into anomaly rankings. The tabular foundation model remains frozen throughout; no parameter update is performed.

3.2 Bootstrap Positive Evidence

For each query fold $\mathcal{Q}_b$, we first ask which rows in its evidence-mining fold $\mathcal{M}_b$ deviate from the predictive structure captured by normal references. To obtain this initial evidence, we randomly select a small set of feature columns $\mathcal{P}$, shared throughout the procedure. For each $j \in \mathcal{P}$, feature j is masked and predicted from the remaining features using $\mathcal{S}_{\text{ref}}$ as context. The frozen model returns both a point prediction and predictive quantiles for the masked feature.

We derive two complementary signals from these predictions. They capture different aspects of deviation from normal predictive structure: one measures how unusual the conditional prediction error is, while the other measures how weakly the observed value is locally supported by the predictive distribution.

For the first signal, let

$$r_j(x) = x_j - \hat{f}_j(x_{-j})$$

denote the conditional prediction residual. We robustly normalize this residual relative to the other rows in $\mathcal{M}_b$:

$$z_j(x; \mathcal{M}_b) = \frac{|r_j(x) - \text{med}_{u \in \mathcal{M}_b} r_j(u)|}{\max(c_{\text{MAD}} \ \text{MAD}_{u \in \mathcal{M}_b} r_j(u), 10^{-9})}.$$

Large z_j indicates an unusually large conditional prediction error within the evidence-mining fold.

The second signal, $\nu_j(x)$, measures the spacing between predictive quantiles surrounding the observed value x_j. Because these quantiles enclose a fixed amount of predictive probability mass, larger spacing indicates weaker local conditional support around the observed value. The exact quantile indexing and boundary handling are given in Appendix C.

Finally, we convert both signals to percentile ranks within $\mathcal{M}_b$ and average them across the selected features:

$$s_0(x; \mathcal{M}_b) = \frac{1}{2|\mathcal{P}|} \sum_{j \in \mathcal{P}} \left[\text{prank}_{\mathcal{M}_b}(z_j(x; \mathcal{M}_b)) + \text{prank}_{\mathcal{M}_b}(\nu_j(x)) \right].$$

The percentile-rank definition and tie handling are provided in Appendix A.

Rows with larger s_0 provide stronger candidate positive evidence. Importantly, s_0 is a *bootstrap evidence score*, not the final anomaly score. The next stage uses this ranking to construct a discriminative in-context task and re-evaluates the evidence against known normal observations.

3.3 SELF-CONSTRUCTED DISCRIMINATIVE REFINEMENT

The bootstrap ranking indicates where positive evidence may lie, but it is not yet a discriminative anomaly ranking. We therefore use the bootstrap evidence to construct an uncertainty-aware *in-context regression* task.

For evidence-mining fold $\mathcal{M}_b$, we initialize

$$c_b^{(0)}(x) = s_0(x; \mathcal{M}_b), \qquad x \in \mathcal{M}_b.$$

More generally, let c denote the current evidence score on a set of target rows $\mathcal{V}$. We select a small set $N \subseteq \mathcal{V}$ of high-evidence rows as positive evidence.

Known normal rows in $\mathcal{S}_{\text{ref}}$ receive target 0, while rows in N receive target 1. Because weak positive evidence does not imply normality, we avoid assigning all remaining target rows to the normal class. Instead, a subset of the remaining rows, denoted $\mathcal{U}(N)$, receives graded intermediate targets in $[0.5, 0.75]$ according to the current evidence ranking. This preserves uncertainty while retaining their relative evidence ordering.

The resulting in-context regression task is

$$\mathcal{C}(N; c, \mathcal{V}) = \Big[(\mathcal{S}_{\text{ref}}, 0),\ (N, 1),\ (\mathcal{U}(N), y_c)\Big],$$

where y_c denotes the graded intermediate targets. Given this context, the same frozen tabular foundation model is queried in regression mode, and its output for a query row x is denoted by

$$d(x; N, c, \mathcal{V}).$$

This serves as the discriminative score rather than a calibrated probability. The exact target mapping, evidence-set sizes, and unresolved-row sampling are provided in Appendix C.

Directly using the same rows both to define positive evidence and to re-score themselves could amplify errors in the initial ranking. To prevent this, each evidence-mining fold is divided into two disjoint inner folds,

$$\mathcal{M}_b = \mathcal{M}_{b,0} \cup \mathcal{M}_{b,1}, \qquad \mathcal{M}_{b,0} \cap \mathcal{M}_{b,1} = \varnothing.$$

At refinement round t, positive evidence selected from one inner fold is used only to score the other. Let $N_{b,\bar{\phi}}^{(t)}$ denote the positive-evidence set selected from $\mathcal{M}_{b,\bar{\phi}}$, where $\bar{\phi} = 1 - \phi$. For $x \in \mathcal{M}_{b,\phi}$,

$$c_b^{(t)}(x) = d\Big(x; N_{b,\bar{\phi}}^{(t)}, c_b^{(t-1)}, \mathcal{M}_{b,\bar{\phi}}\Big).$$

Both directions use the previous-round scores and are updated synchronously.

After R rounds, $c_b^{(R)}$ defines the refined evidence ranking used for final scoring. Cross-fitting prevents a row from directly reinforcing its own provisional role: evidence can influence a row only through the scores it induces on other rows.

3.4 EVIDENCE RESAMPLING AND FINAL SCORING

Even after refinement, it remains uncertain which target rows should serve as positive evidence. Relying on a single top-k set would make the final prediction sensitive to where this cutoff is placed. We therefore sample multiple positive-evidence sets from the refined ranking and average their predictions.

For each evidence-mining fold $\mathcal{M}_b$, we convert the refined scores $c_b^{(R)}$ into sampling probabilities,

$$p_b(x) \propto \exp\left(\frac{\text{prank}_{\mathcal{M}_b}(c_b^{(R)}(x)) - 1}{\tau}\right), \qquad x \in \mathcal{M}_b,$$

where τ controls how strongly sampling favors high-evidence rows. We then draw K positive-evidence sets

$$N_{b,k} \sim \text{WSample}\left(\mathcal{M}_b, k_{\text{fin}}(\mathcal{M}_b), p_b\right), \qquad k = 1, \ldots, K,$$

without replacement within each draw. Thus, rows with stronger refined evidence are more likely to serve as positive context, while uncertainty around any single cutoff is retained across multiple contexts. The exact sampling rule and default hyperparameters are given in Appendix C.

Each sampled set defines a discriminative context

$$\mathcal{C}_{b,k} = \mathcal{C}\left(N_{b,k}; c_b^{(R)}, \mathcal{M}_b\right),$$

and the frozen tabular foundation model produces the corresponding discriminative score

$$d_{b,k}(x) = d\Big(x; N_{b,k}, c_b^{(R)}, \mathcal{M}_b\Big).$$

Because different evidence contexts can produce scores on different scales, we align each context using the held-out normal calibration set $\mathcal{S}_{\text{cal}}$. The calibration scores

$$\{d_{b,k}(u) : u \in \mathcal{S}_{\text{cal}}\}$$

define a monotone reference-based alignment map $A_{b,k}$. Its exact construction, including the treatment of scores outside the observed calibration range, is given in Appendix E.

The final anomaly score for $x \in \mathcal{Q}_b$ is

$$g_b(x) = \frac{1}{K}\sum_{k=1}^{K} A_{b,k}(d_{b,k}(x)).$$

In this way, SCOPE averages over uncertainty in which observations should serve as positive evidence rather than committing to a single hard evidence set. Outer cross-fitting ensures that a row in $\mathcal{Q}_b$ never contributes to the evidence used to score itself, and the foundation model remains frozen throughout.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate SCOPE on three large-scale tabular anomaly-detection benchmarks: ADBench, OddBench, and OvRBench. Our evaluation contains 57, 690, and 755 benchmark entries, respectively, for a total of 1,502 entries. We use ADBench for component analysis and apply the resulting SCOPE configuration unchanged to OddBench and OvRBench. We compare against classical, deep, and foundation-model-based anomaly detectors using the per-dataset results released with the corresponding benchmarks. Methods unavailable in a benchmark's released records are left blank. Further details on dataset construction, baseline sources, and reproduction are provided in Appendix K.

Evaluation. We evaluate predictive performance using AUROC and AUPRC. To summarize performance across heterogeneous datasets, we additionally report several rank- and pairwise-based aggregate statistics, defined in Table 1. Because these statistics depend on the comparison set, we recompute them after adding SCOPE rather than reusing published aggregate values. Before adding SCOPE, our recomputation closely matches the benchmark authors' published aggregates, with seven of eight checks agreeing to within 5×10^{-4} and the remaining deviation below 9×10^{-4} (Appendix K).

4.2 Main Results

Tables 1 and 2 report the main comparison. Across AUROC and AUPRC, SCOPE ranks first under all five aggregate metrics on all three benchmarks, yielding the best result in all 30 benchmark–metric combinations.

For AUROC, SCOPE attains average ranks of 3.37, 3.53, and 2.33 on ADBench, OddBench, and OvRBench, respectively. The corresponding second-best values are 4.77, 6.19, and 6.51. The advantage is also reflected in pairwise win rates of .798, .798, and .893. Moreover, SCOPE obtains

Table 1: **Main results: AUROC.** R = average rank, Elo = Elo averaged over 100 random dataset orderings, W = pairwise win rate, rAUC = rescaled AUC, and CΔ = champion delta, the mean relative error gap to the best method on each dataset. Lower is better for R and CΔ; higher is better otherwise. Best values are bold and second-best values are underlined.

	ADBench (57)					OddBench (690)					OvRBench (755)				
Method	R↓	Elo↑	W↑	rAUC↑	CΔ ↓	R↓	Elo↑	W↑	rAUC↑	CΔ ↓	R↓	Elo↑	W↑	rAUC↑	CΔ ↓
Classical															
kNN	6.05	1083	.558	.907	.538	6.86	1053	.551	.848	.555	6.87	1054	.567	.862	.480
LOF	6.81	1015	.504	.884	.563	7.95	974	.481	.818	.608	7.93	991	.493	.829	.539
IForest	9.25	868	.311	.851	.636	8.23	1021	.482	.834	.607	8.18	991	.484	.850	.527
OCSVM	–	–	–	–	–	8.25	979	.467	.841	.625	8.61	950	.446	.845	.550
CBLOF	–	–	–	–	–	8.53	992	.438	.824	.599	8.48	950	.455	.842	.528
EGMM	–	–	–	–	–	6.44	1080	.581	.839	.507	6.51	1081	.590	.861	.468
Deep															
DeepSVDD	10.77	722	.186	.769	.717	–	–	–	–	–	–	–	–	–	–
GOAD	10.21	749	.219	.779	.663	12.21	688	.184	.632	.708	12.43	706	.178	.666	.651
ICL	6.40	1027	.548	.895	.543	9.54	952	.381	.798	.654	7.68†	1002	.440	.841	.526
NPT-AD	–	–	–	–	–	11.21	767	.263	.710	.674	11.50	737	.242	.738	.609
DDPM	7.86	941	.411	.875	.585	–	–	–	–	–	–	–	–	–	–
DTE-C	6.81	1026	.504	.902	.562	7.91	1005	.483	.830	.589	7.68†	1002	.440	.841	.526
DTE-NP	5.33	1104	.624	.915	.542	6.19	1104	.607	.859	.548	6.71	1074	.580	.861	.496
Foundation-model-based															
FoMo-0D	6.95	1019	.500	.899	.595	7.81	963	.490	.827	.580	8.65	934	.443	.823	.555
TabPFN-OD	5.65	1087	.601	.910	.536	6.83	1047	.555	.854	.543	7.26	1017	.539	.858	.479
OutFormer	4.77	1121	.670	.925	.498	6.36	1090	.585	.858	.540	7.03	1078	.556	.858	.497
SCOPE	**3.37**	**1238**	**.798**	**.962**	**.188**	**3.53**	**1285**	**.798**	**.948**	**.197**	**2.33**	**1433**	**.893**	**.978**	**.084**

†The released OvRBench records contain byte-identical ICL and DTE-C columns. Removing the duplicate does not change the conclusions. Ranking statistics are recomputed from the released per-dataset records with SCOPE included; reproduction details are given in Appendix K.

Table 2: **Main results: AUPRC.** Metrics and notation follow Table 1. Best values are bold and second-best values are underlined.

	ADBench (57)					OddBench (690)					OvRBench (755)				
Method	R↓	Elo↑	W↑	rAUC↑	CΔ ↓	R↓	Elo↑	W↑	rAUC↑	CΔ ↓	R↓	Elo↑	W↑	rAUC↑	CΔ ↓
Classical															
kNN	6.40	1053	.538	.754	.497	7.52	1025	.508	.613	.442	7.37	1007	.533	.707	.443
LOF	6.95	994	.493	.715	.532	7.62	991	.509	.607	.461	7.86	1013	.500	.679	.468
IForest	9.79	792	.266	.600	.581	9.22	935	.410	.525	.500	8.87	916	.434	.648	.489
OCSVM	–	–	–	–	–	7.16	1027	.547	.624	.470	7.84	1000	.502	.692	.474
CBLOF	–	–	–	–	–	8.74	972	.428	.569	.465	8.69	925	.442	.665	.470
EGMM	–	–	–	–	–	6.53	1096	.578	.656	.399	6.41	1086	.601	.729	.422
Deep															
DeepSVDD	9.96	769	.249	.606	.605	–	–	–	–	–	–	–	–	–	–
GOAD	10.09	782	.231	.602	.562	11.84	731	.212	.433	.508	12.20	748	.196	.514	.535
ICL	6.16	1053	.569	.740	.496	8.91	986	.426	.575	.484	8.17†	978	.500	.675	.470
NPT-AD	–	–	–	–	–	11.25	767	.258	.450	.505	11.23	772	.262	.547	.518
DDPM	7.70	968	.431	.712	.544	–	–	–	–	–	–	–	–	–	–
DTE-C	7.21	995	.472	.714	.532	8.43	967	.448	.585	.455	8.17†	978	.407	.675	.470
DTE-NP	5.58	1085	.607	.768	.505	6.67	1073	.574	.635	.441	6.81	1064	.575	.716	.444
Foundation-model-based															
FoMo-0D	6.65	1040	.529	.758	.523	7.25	1016	.533	.649	.429	8.06	983	.487	.683	.467
TabPFN-OD	5.61	1080	.602	.744	.499	6.86	1047	.556	.643	.418	7.08	1022	.554	.714	.424
OutFormer	5.25	1110	.633	.794	.495	6.18	1084	.605	.662	.409	6.75	1089	.578	.719	.430
SCOPE	**3.04**	**1279**	**.829**	**.915**	**.111**	**3.96**	**1282**	**.773**	**.814**	**.105**	**2.55**	**1419**	**.881**	**.942**	**.061**

substantially smaller champion deltas (.188, .197, and .084) than the next-best values (.498, .507, and .468), indicating that its performance remains close to the dataset-specific best method across heterogeneous tasks.

The same pattern holds for AUPRC. SCOPE obtains average ranks of 3.04, 3.96, and 2.55, compared with second-best values of 5.25, 6.18, and 6.41. Its pairwise win rates are .829, .773, and .881, while its champion deltas are .111, .105, and .061, respectively. The particularly strong OvRBench results—average ranks of 2.33 for AUROC and 2.55 for AUPRC, with win rates of .893 and .881—show that the advantage extends across a large collection of one-vs-rest anomaly-detection tasks. These results support the central premise of SCOPE: a general-purpose frozen tabular foundation model can be turned into a strong anomaly detector by constructing discriminative supervision from the available data, without anomaly-specific pretraining or parameter updates.

4.3 Mechanism and Component Analysis

Where do the gains come from? The design of SCOPE separates the signal used to locate candidate positive evidence from the discriminative score used for final ranking. Table 3 follows this transition from the bootstrap score s_0, through its first conversion into a discriminative in-context task, to iterative cross-fitted refinement. The largest transition occurs before iterative refinement. Replacing s_0 as the final score with the constructed discriminative task improves AUROC by 7.79 points and AUPRC by 18.73 points ($p = 7.2 \times 10^{-6}$ and 7.0×10^{-8}, respectively). Thus, the predictive discrepancies used to bootstrap SCOPE are most useful for locating possible positive evidence; the final ranking is substantially improved when that evidence is turned into an in-context discrimination problem. Cross-fitted refinement then provides a smaller but statistically clear correction. At the default $R = 3$, performance improves by a further 0.50 AUROC and 1.46 AUPRC over $R = 0$ ($p = 0.011$ and 2.4×10^{-4}). The gain largely appears after the first round, and performance is subsequently non-monotonic but similar across $R = 1$, 3, and 5. We therefore use $R = 3$ as a single fixed setting rather than tuning the number of refinement rounds for individual datasets.

Table 3: **Stage decomposition** on ADBench.

Stage	AUROC	AUPRC
Bootstrap s_0	77.10	47.23
Discriminative ($R = 0$)	84.89	65.96
$R = 1$	85.36	**68.24**
$R = 3$ (default)	**85.39**	67.43
$R = 5$	85.34	68.00

Core component ablations. Table 4 examines the main choices involved in constructing and using this discriminative task: preventing self-reinforcement during refinement, representing unresolved target rows, accounting for uncertainty in the selected positive evidence, aligning scores across evidence contexts, and choosing the prediction interface. Full paired statistics and additional variants are reported in Appendix G.

Cross-fitting has the largest effect. When refinement is retained but the fold separation is removed, performance drops by 5.27 AUROC and 4.91 AUPRC. This ablation is distinct from $R = 0$ above: $R = 0$ omits refinement altogether, whereas this variant performs refinement while allowing provisional evidence to influence scores on the same rows from which that evidence was selected. The large degradation therefore supports cross-fitting as a mechanism for limiting self-reinforcement during evidence refinement. The unresolved-target ablations separate whether uncertain rows should be represented at all from how they should be encoded. Excluding unresolved rows and retaining only the binary normal/positive-evidence context reduces performance by 1.11 AUROC and 2.41 AUPRC. Including them with a constant target of 0.5 recovers most of this loss, while the full SCOPE uses graded targets in $[0.5, 0.75]$ to additionally preserve their relative evidence ordering. The graded encoding provides a further 0.37 AUROC and 0.41 AUPRC improvement.

Table 4: **Core ablations** on ADBench. Differences are from full SCOPE.

Variant	ΔROC	ΔPR
w/o cross-fitting	−5.27	−4.91
Unresolved-target encoding		
excluded (binary)	−1.11	−2.41
constant 0.5	−0.37	−0.41
single context ($K = 1$)	−0.25	−0.93
w/o tail extension	−1.11	−4.03
classification interface	−0.84	−2.27

The remaining ablations concern how refined evidence is used at inference. Using a single sampled evidence context ($K = 1$) decreases performance by 0.25 AUROC and 0.93 AUPRC, supporting aggregation over multiple plausible positive-evidence sets rather than reliance on one sampled set. Removing the reference-tail extension causes a larger loss, especially in AUPRC (−4.03): without the extension, all query scores above the largest reference score receive the same empirical-CDF value, eliminating resolution in the upper tail.

Finally, replacing the regression interface with the backbone's classification interface decreases performance by 0.84 AUROC and 2.27 AUPRC. This is consistent with the role of regression in representing unresolved rows through intermediate targets rather than forcing the constructed task to be binary. Together, these results show that the full performance of SCOPE reflects contributions from several distinct design choices rather than a single scoring heuristic.

Across Frozen Tabular Backbones We use TabICL (Qu et al., 2026) as the default backbone because its in-context regression interface directly supports the graded targets used by SCOPE, and we keep this backbone fixed throughout the main experiments to avoid conflating method effects with backbone selection. As a separate backbone analysis, we replace TabICL with EXAONE-Tabular (Eo et al., 2026) and TabPFN-2.5 (Grinsztajn et al., 2025) while leaving the rest of the pipeline unchanged. EXAONE-Tabular is numerically higher but not significantly different from TabICL, whereas TabPFN-2.5 performs lower. Thus, the gains of SCOPE are not tied to TabICL, although backbone choice still affects final performance.

Table 5: **Backbone comparison** on ADBench.

Backbone	AUROC	AUPRC
TabICL	85.39	67.43
EXAONE-Tabular	**86.45**	**68.35**
TabPFN-2.5	84.51	65.38

Evidence re-ranking. The discriminative stage does not simply preserve the bootstrap evidence. The initial and final top-q sets overlap by only 34.2%, while mean top-q precision increases from 50.4% to 68.8%. Thus, cross-fitted discrimination substantially re-ranks the candidate evidence rather than merely amplifying the bootstrap ordering. Detailed retention, demotion, and rank-correlation statistics are reported in Appendix G.

Stability and robustness. The final scoring procedure is stable to several implementation choices. With fixed evidence contexts and reference-based alignment, shuffling the query order leaves scores exactly unchanged on all diagnostic datasets. Changing query batch size produces nearly identical scores on five of six datasets; the remaining dataset shows larger score shifts but nearly unchanged AUROC, indicating that this residual variation arises from the frozen backbone rather than the alignment map itself. Hyperparameter sensitivity is similarly mild around the fixed default configuration: reducing the evidence fraction from $q = 0.05$ to 0.02 produces essentially unchanged performance, while increasing the normal-reference budget from $B_{\text{ctx}} = 2048$ to 4096 yields no clear improvement. Full batch diagnostics and sensitivity results for q, the number of feature probes P, and B_{ctx} are reported in Appendices H and G.

5 Discussion and Conclusion

Computational cost. SCOPE requires no gradient-based optimization or parameter update; all computation consists of frozen-model inference. On a single idle A100, the default configuration takes 51.5 s per dataset on average over eight measured ADBench datasets (523–5250 test rows, $d = 6$–768). Once the final evidence contexts are fixed, subsequent batched prediction requires only the $K = 4$ final scoring calls. Detailed stage-wise and scaling analyses are provided in Appendix L.

Verified feedback under extreme anomaly rarity. Extremely low anomaly prevalence is the most challenging regime for self-constructed positive evidence. On the 53 OddBench datasets with contamination below 1%, however, SCOPE can directly incorporate verified anomalies as positive in-context evidence without retraining or parameter updates. A single verified anomaly improves AUROC by 1.91 and AUPRC by 2.76; with two, the gains increase to $+3.35$ and $+5.76$. Notably, one verified anomaly is already sufficient to move SCOPE from behind the strongest published AUROC baseline in average rank (7.04 vs. 6.58) to ahead of it (6.27 vs. 6.64). This provides a natural path from fully label-free detection to lightweight human-in-the-loop adaptation as verified anomalies become available. Full results are given in Appendix M.

Conclusion. We introduced SCOPE, which turns predictive signals from a frozen general-purpose tabular foundation model into positive evidence for in-context anomaly discrimination. Across three large-scale benchmarks, this construction consistently improves over strong classical, deep, and foundation-model-based detectors. Our analyses show that most of the gain arises from converting bootstrapped predictive evidence into a discriminative task, with cross-fitted refinement providing a further correction. More broadly, the results suggest that anomaly detection need not require anomaly-specific pretraining: useful supervision can instead be constructed at inference time from the target data itself.

AI USE STATEMENT

We used generative AI tools for translation, language editing, and improving the clarity and presentation of the manuscript. All AI-assisted text and suggestions were reviewed by the authors. The authors take full responsibility for the final content of this paper, including all claims, experimental results, and conclusions.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. The study does not involve human participants or animal experiments and uses publicly available tabular anomaly-detection benchmarks. We are not aware of any additional privacy, security, or ethical concerns arising from the experiments conducted in this work.

REPRODUCIBILITY STATEMENT

We provide the source code in the supplementary material. Detailed implementation settings for bootstrap evidence construction, cross-fitted refinement, evidence resampling, score alignment, and all hyperparameters are described in the main text and appendices. Benchmark construction, baseline sources, and evaluation procedures are also documented, and SCOPE uses a fixed configuration across benchmarks unless otherwise stated.

REFERENCES

Lion Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *International Conference on Management of Data*, 2000.

Xueying Ding, Simon Klüttermann, Haomin Wen, Yilong Chen, and Leman Akoglu. Macrodata: New benchmarks of thousands of datasets for tabular outlier detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026a.

Xueying Ding, Haomin Wen, Simon Klüttermann, and Leman Akoglu. From zero to hero: Advancing zero-shot foundation models for tabular outlier detection. In *International Conference on Machine Learning (ICML)*, 2026b.

Moonjung Eo, Min-Kook Suh, Hye-Seung Cho, Jiwon Kim, Seoyoon Kim, Sangjun Nam, and Soonyoung Lee. EXAONE Tabular 1.0: Technical report. *arXiv preprint*, 2026.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International Conference on Machine Learning (ICML)*, 2020. URL https://arxiv.org/abs/2002.12718.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, Mihir Manium, Rosen Yu, Felix Jablonski, Shi Bin Hoo, Anurag Garg, Jake Robertson, Magnus Bühler, Vladyslav Moroshan, Lennart Purucker, Clara Cornu, Lilly Charlotte Wehrhahn, Alessandro Bonetto, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-2.5: Advancing the state of the art in tabular foundation models, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, 2022.

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

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019.

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *ICLR*, 2024.

Sha Lu, Jixue Liu, Stefan Peters, Thuc Duy Le, Craig Xie, Lin Liu, and Jiuyong Li. ulead-tabpfn: Uncertainty-aware dependency-based anomaly detection with tabpfn. *arXiv preprint*, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. In *International Conference on Machine Learning (ICML)*, 2026.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pp. 427–438, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335437.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pp. 4393–4402, 2018.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf.

Yuchen Shen, Haomin Wen, and Leman Akoglu. FoMo-0D: A foundation model for zero-shot tabular outlier detection. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*, 2022.

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. Beyond individual input for deep anomaly detection on tabular data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48097–48123. PMLR, 21–27 Jul 2024.

Jack Yi Wei and Narges Armanfard. Iclad: In-context learning for unified tabular anomaly detection across supervision regimes. *arXiv preprint*, 2026.

Suhee Yoon, Sanghyu Yoon, Ye Seul Sim, Seungdong Yoa, Dongmin Kim, Soonyoung Lee, Hankook Lee, and Woohyung Lim. AutoAnoEval: Semantic-aware model selection via tree-guided LLM reasoning for tabular anomaly detection. In *Findings of the Association for Computational Linguistics: EACL 2026*. Association for Computational Linguistics, 2026.

Table 6: Notation used by SCOPE.

Symbol	Meaning
$\mathcal{S}_{\text{all}}$	Available normal training observations.
$\mathcal{S}$	Normal pool sampled from $\mathcal{S}_{\text{all}}$.
$\mathcal{S}_{\text{ref}}$	Normal reference rows used for conditional feature prediction and discriminative in-context regression.
$\mathcal{S}_{\text{cal}}$	Held-out normal rows used for final score alignment.
$\mathcal{Q}$	Entire unlabeled target set to be scored.
$\mathcal{Q}_b$	Outer query fold being scored, where $b \in \{A, B\}$.
$\mathcal{M}_b$	Evidence-mining fold used to construct evidence for $\mathcal{Q}_b$; $\mathcal{M}_A = \mathcal{Q}_B$ and $\mathcal{M}_B = \mathcal{Q}_A$.
$\mathcal{M}_{b,\phi}$	Inner fold of $\mathcal{M}_b$, where $\phi \in \{0, 1\}$.
$\mathcal{P}$	Randomly selected feature set used for conditional feature prediction.
$\mathcal{V}$	Generic target-row set from which a discriminative context is constructed.
$\hat{f}_j(x_{-j})$	Point prediction of feature j from the remaining features using $\mathcal{S}_{\text{ref}}$ as context.
$\hat{Q}_{j,\ell}(x_{-j})$	Predictive quantile of feature j at level α_ℓ.
$r_j(x)$	Conditional prediction residual for feature j.
$z_j(x; \mathcal{V})$	Robust residual signal relative to rows in $\mathcal{V}$.
$\nu_j(x)$	Predictive-quantile spacing around the observed value of feature j.
$\text{prank}_{\mathcal{V}}$	Percentile rank computed among rows in $\mathcal{V}$.
$s_0(x; \mathcal{V})$	Bootstrap evidence score.
$c_b^{(t)}(x)$	Evidence score after refinement round t; $c_b^{(0)}$ denotes the bootstrap evidence score.
N	Rows selected from $\mathcal{V}$ to serve as positive evidence.
$\mathcal{U}(N)$	Additional rows from $\mathcal{V} \setminus N$ included with graded intermediate targets.
$y_c(x)$	Graded regression target assigned to a row in $\mathcal{U}(N)$.
$\mathcal{C}(N; c, \mathcal{V})$	Discriminative in-context regression context constructed from known normals, positive evidence, and uncertain target rows.
$d(x; N, c, \mathcal{V})$	Regression output of the frozen tabular foundation model for query row x under $\mathcal{C}(N; c, \mathcal{V})$; used as a discriminative score.
$p_b(x)$	Probability with which $x \in \mathcal{M}_b$ is sampled as positive evidence during final evidence resampling.
$N_{b,k}$	k-th sampled positive-evidence set for query fold $\mathcal{Q}_b$.
$d_{b,k}(x)$	Raw discriminative score under the context defined by $N_{b,k}$.
$A_{b,k}$	Reference-based monotone alignment map for context (b, k).
$g_b(x)$	Final anomaly score for $x \in \mathcal{Q}_b$.

A Notation

Table 6 summarizes the notation used by SCOPE. We use $\mathcal{V}$ only as a generic target-row set when defining reusable operations; in the cross-fitted procedure, evidence for query fold $\mathcal{Q}_b$ is constructed from the opposite evidence-mining fold $\mathcal{M}_b$.

B Full SCOPE Algorithm

Algorithm 1 summarizes the complete procedure. The normal pool, feature set, and all outer and inner partitions are sampled once per run and then held fixed. All prediction calls use the same frozen tabular foundation model; no parameter update is performed.

C Implementation Details

Normal pool and calibration reserve. We uniformly sample a normal pool

$$|\mathcal{S}| = \min\left(B_{\text{ctx}}, |\mathcal{S}_{\text{all}}|\right), \qquad B_{\text{ctx}} = 2048.$$

The calibration reserve has size

$$|\mathcal{S}_{\text{cal}}| = \min\left(B_{\text{cal}}, \lfloor |\mathcal{S}|/5 \rfloor\right), \qquad B_{\text{cal}} = 256,$$

Algorithm 1: SCOPE: Self-Constructed Positive Evidence

Require: Normal observations $\mathcal{S}_{\text{all}}$, unlabeled target set $\mathcal{Q}$

Sample normal pool $\mathcal{S} \subseteq \mathcal{S}_{\text{all}}$ and divide it into $\mathcal{S}_{\text{ref}}$ and $\mathcal{S}_{\text{cal}}$
Sample shared feature set $\mathcal{P}$
Split $\mathcal{Q}$ into disjoint outer folds $\mathcal{Q}_A$ and $\mathcal{Q}_B$
Set $\mathcal{M}_A \leftarrow \mathcal{Q}_B$ and $\mathcal{M}_B \leftarrow \mathcal{Q}_A$
for $b \in \{A, B\}$ **do**
    Compute bootstrap evidence scores $c_b^{(0)}(x) \leftarrow s_0(x; \mathcal{M}_b)$ for all $x \in \mathcal{M}_b$
    Split $\mathcal{M}_b$ into disjoint inner folds $\mathcal{M}_{b,0}$ and $\mathcal{M}_{b,1}$
    for $t = 1, \ldots, R$ **do**
        for $\phi \in \{0, 1\}$ **do**
            Set $\bar{\phi} \leftarrow 1 - \phi$
            Select positive evidence $N_{b,\bar{\phi}}^{(t)} \leftarrow \text{TopK}\left(c_b^{(t-1)}; \mathcal{M}_{b,\bar{\phi}}, k_{\text{ref}}(\mathcal{M}_{b,\bar{\phi}})\right)$
            Construct $\mathcal{C}\left(N_{b,\bar{\phi}}^{(t)}; c_b^{(t-1)}, \mathcal{M}_{b,\bar{\phi}}\right)$
            For all $x \in \mathcal{M}_{b,\phi}$, set $c_b^{(t)}(x) \leftarrow d\left(x; N_{b,\bar{\phi}}^{(t)}, c_b^{(t-1)}, \mathcal{M}_{b,\bar{\phi}}\right)$
        end for
        Commit the two inner-fold updates synchronously
    end for
    Convert $c_b^{(R)}$ into evidence-sampling probabilities p_b
    for $k = 1, \ldots, K$ **do**
        Sample $N_{b,k} \sim \text{WSample}\left(\mathcal{M}_b, k_{\text{fin}}(\mathcal{M}_b), p_b\right)$
        Construct $\mathcal{C}_{b,k} \leftarrow \mathcal{C}\left(N_{b,k}; c_b^{(R)}, \mathcal{M}_b\right)$
        For all $x \in \mathcal{S}_{\text{cal}} \cup \mathcal{Q}_b$, compute $d_{b,k}(x)$ under $\mathcal{C}_{b,k}$
        Construct $A_{b,k}$ from $\{d_{b,k}(u) : u \in \mathcal{S}_{\text{cal}}\}$
    end for
    For all $x \in \mathcal{Q}_b$, set $g_b(x) \leftarrow K^{-1} \sum_{k=1}^{K} A_{b,k}(d_{b,k}(x))$
end for
return Cross-fitted anomaly scores on $\mathcal{Q}$

and the remaining rows form $\mathcal{S}_{\text{ref}}$.

The reserve requires at least 32 rows. If fewer than 32 rows can be reserved, the disjoint split is skipped and the calibration reference is sampled from within $\mathcal{S}_{\text{ref}}$. Among the datasets considered in our runs, only `Parkinson` ($|\mathcal{S}| = 114$) falls below this threshold, and it does not enter the reported paired comparisons because it is absent from the released reference records. Thus, all 1,502 benchmark entries reported in the main comparison use the disjoint calibration reserve.

Importantly, B_{ctx} limits only the normal pool and is not a bound on the total discriminative context size.

Percentile ranks. For a scalar statistic a evaluated on a finite set $\mathcal{V}$, we define

$$\text{prank}_{\mathcal{V}}(a(x)) = \frac{\text{rank}^{\text{avg}}\left(a(x); \{a(v) : v \in \mathcal{V}\}\right)}{|\mathcal{V}|}.$$

Ranks are ascending and ties receive their average rank. A unique maximum therefore receives exactly 1, while tied maxima share their average rank.

Selected feature set. We sample

$$|\mathcal{P}| = \min(P, d), \qquad P = 3,$$

feature columns uniformly without replacement once per run. The same feature set $\mathcal{P}$ is shared across both outer folds.

Conditional prediction calls. For each $j \in \mathcal{P}$, feature j is masked and predicted from the remaining features using $\mathcal{S}_{\text{ref}}$ as context. The point prediction $\hat{f}_j(x_{-j})$ and predictive quantiles $\hat{Q}_{j,\ell}(x_{-j})$ are obtained through separate prediction calls to the same frozen tabular foundation model.

Robust residual signal. For feature j,

$$r_j(x) = x_j - \hat{f}_j(x_{-j}).$$

For a target-row set $\mathcal{V}$, let

$$m_j^{\mathcal{V}} = \text{median}_{u \in \mathcal{V}}\, r_j(u),$$

and

$$\text{MAD}_j^{\mathcal{V}} = \text{median}_{u \in \mathcal{V}} \left| r_j(u) - m_j^{\mathcal{V}} \right|.$$

The normalized residual signal is

$$z_j(x; \mathcal{V}) = \frac{|r_j(x) - m_j^{\mathcal{V}}|}{\max\left(c_{\text{MAD}}\, \text{MAD}_j^{\mathcal{V}}, 10^{-9}\right)}, \qquad c_{\text{MAD}} = \left[\Phi^{-1}(0.75)\right]^{-1},$$

where Φ is the standard normal cumulative distribution function.

In the bootstrap stage, $\mathcal{V} = \mathcal{M}_b$, so the residual center and scale are estimated only from the corresponding evidence-mining fold.

Predictive-quantile spacing. We use $L = 19$ predictive quantiles at

$$\alpha_\ell = \frac{\ell}{20}, \qquad \ell = 1, \ldots, 19.$$

Let

$$\hat{Q}_{j,1}(x_{-j}) \le \cdots \le \hat{Q}_{j,L}(x_{-j})$$

denote the sorted predictive quantiles.

For

$$\text{clip}(a, l, u) = \min\{\max\{a, l\}, u\},$$

we first locate the observed value relative to these quantiles:

$$u_j(x) = \text{clip}\left(\frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}\left\{\hat{Q}_{j,\ell}(x_{-j}) < x_j\right\}, 10^{-6}, 1 - 10^{-6}\right).$$

We then define

$$c_j(x) = 1 + \text{clip}\left(\lfloor u_j(x)(L-1) \rfloor, 1, L-2\right) \in \{2, \ldots, L-1\}.$$

The spacing signal is

$$\nu_j(x) = \log \max\left(\hat{Q}_{j,c_j(x)+1}(x_{-j}) - \hat{Q}_{j,c_j(x)-1}(x_{-j}), 10^{-12}\right).$$

Queries outside the predicted quantile range use the outermost available centered interval. Because the two surrounding quantiles cover a fixed amount of predictive probability mass, larger spacing indicates weaker local conditional support. We do not interpret ν_j as a density estimate or as a measure of how far an observation extends beyond the predictive tails.

Bootstrap evidence score. For a target-row set $\mathcal{V}$,

$$s_0(x; \mathcal{V}) = \frac{1}{2|\mathcal{P}|} \sum_{j \in \mathcal{P}} \left[\text{prank}_{\mathcal{V}}\left(z_j(x; \mathcal{V})\right) + \text{prank}_{\mathcal{V}}\left(\nu_j(x)\right)\right].$$

In the cross-fitted procedure,

$$c_b^{(0)}(x) = s_0(x; \mathcal{M}_b), \qquad x \in \mathcal{M}_b.$$

Evidence-set size. For a target-row set $\mathcal{V}$, define

$$k_\bullet(\mathcal{V}) = \min\left(|\mathcal{V}|, B_{\max}, \max\left(k_\bullet^{\min}, \mathrm{round}(q|\mathcal{V}|)\right)\right),$$

where $\bullet \in \{\mathrm{ref}, \mathrm{fin}\}$. We use

$$q = 0.05, \qquad B_{\max} = 1024, \qquad k_{\mathrm{ref}}^{\min} = 2, \qquad k_{\mathrm{fin}}^{\min} = 8.$$

During refinement, $\mathrm{TopK}(c; \mathcal{V}, k)$ returns the k rows in $\mathcal{V}$ with the largest current evidence scores under c.

Graded targets for uncertain rows. For selected positive evidence $N \subseteq \mathcal{V}$, the remaining rows are neither discarded nor treated as normal. We define

$$\mathcal{U}(N) \subseteq \mathcal{V} \setminus N, \qquad |\mathcal{U}(N)| \le B_{\max}.$$

If $|\mathcal{V} \setminus N| \le B_{\max}$, all remaining rows are used. Otherwise, $B_{\max}$ rows are sampled uniformly without replacement from $\mathcal{V} \setminus N$.

Their regression targets are

$$y_c(x) = \mu + \beta\left[2\,\mathrm{prank}_{\mathcal{U}(N)}\left(c(x)\right) - 1\right],$$

where

$$\mu = 0.625, \qquad \beta = 0.125.$$

Hence

$$y_c(x) \in [0.5, 0.75].$$

The three target regimes are therefore

$$0 = \text{known normal}, \qquad [0.5, 0.75] = \text{uncertain target}, \qquad 1 = \text{positive evidence}.$$

The complete discriminative context is

$$\mathcal{C}(N; c, \mathcal{V}) = \left(\begin{bmatrix} \mathcal{S}_{\mathrm{ref}} \\ N \\ \mathcal{U}(N) \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \\ y_c \end{bmatrix}\right).$$

Whenever a context is instantiated, $\mathcal{U}(N)$ is sampled once and then held fixed for all rows scored under that context. We write

$$d(x; N, c, \mathcal{V}) = \mathrm{TFM}_{\mathrm{reg}}\left(x \mid \mathcal{C}(N; c, \mathcal{V})\right)$$

for the regression output of the frozen tabular foundation model. This quantity is used as a discriminative score and is not interpreted as a calibrated probability.

Both the positive-evidence and uncertain-target portions are capped by $B_{\max}$, so

$$|\mathcal{C}(N; c, \mathcal{V})| \le |\mathcal{S}_{\mathrm{ref}}| + 2B_{\max}.$$

Thus, B_{ctx} limits only the normal-reference portion of the context.

Nested cross-fitting. The outer and inner partitions are approximately equal random splits sampled once per run and held fixed across refinement rounds.

For each evidence-mining fold,

$$\mathcal{M}_b = \mathcal{M}_{b,0} \cup \mathcal{M}_{b,1}, \qquad \mathcal{M}_{b,0} \cap \mathcal{M}_{b,1} = \varnothing.$$

At refinement round t, writing $\bar{\phi} = 1 - \phi$, positive evidence is selected only from the opposite inner fold:

$$N_{b,\bar{\phi}}^{(t)} = \mathrm{TopK}\left(c_b^{(t-1)}; \mathcal{M}_{b,\bar{\phi}}, k_{\mathrm{ref}}\left(\mathcal{M}_{b,\bar{\phi}}\right)\right).$$

Rows in $\mathcal{M}_{b,\phi}$ are then scored as

$$c_b^{(t)}(x) = d\left(x; N_{b,\bar{\phi}}^{(t)}, c_b^{(t-1)}, \mathcal{M}_{b,\bar{\phi}}\right), \qquad x \in \mathcal{M}_{b,\phi}.$$

Both inner-fold directions read the same previous-round scores $c_b^{(t-1)}$ and write to disjoint row sets. The refreshed score is formed only after both directions have completed. We use

$$R = 3.$$

Final evidence resampling. The default configuration uses

$$K = 4, \qquad \tau = 0.05.$$

D SCORE-WEIGHTED EVIDENCE RESAMPLING

After refinement, define

$$\rho_b(x) = \mathrm{prank}_{\mathcal{M}_b}\left(c_b^{(R)}(x)\right), \qquad x \in \mathcal{M}_b.$$

The normalized sampling probability is

$$p_b(x) = \frac{\exp\left((\rho_b(x) - 1)/\tau\right)}{\sum_{u \in \mathcal{M}_b} \exp\left((\rho_b(u) - 1)/\tau\right)}.$$

For $k = 1, \ldots, K$, we draw

$$N_{b,k} \sim \mathrm{WSample}\left(\mathcal{M}_b, k_{\mathrm{fin}}(\mathcal{M}_b), p_b\right),$$

where $\mathrm{WSample}(\mathcal{V}, k, p)$ denotes weighted sampling of k rows from $\mathcal{V}$ without replacement according to weights p.

The sampling population is the *entire* evidence-mining fold $\mathcal{M}_b$, rather than a deterministic top-q subset. The refined ranking therefore determines how likely each observation is to serve as positive evidence without imposing a single hard cutoff.

Each sampled set defines

$$\mathcal{C}_{b,k} = \mathcal{C}\left(N_{b,k}; c_b^{(R)}, \mathcal{M}_b\right),$$

and the corresponding raw discriminative score is

$$d_{b,k}(x) = d\left(x; N_{b,k}, c_b^{(R)}, \mathcal{M}_b\right).$$

E REFERENCE-BASED SCORE ALIGNMENT

Different positive-evidence contexts can produce raw discriminative scores on different scales. For each final context (b, k), define the held-out normal-reference scores

$$\mathcal{R}_{b,k} = \{d_{b,k}(u) : u \in \mathcal{S}_{\mathrm{cal}}\}, \qquad n_{\mathrm{cal}} = |\mathcal{S}_{\mathrm{cal}}|.$$

Let

$$t_{\min}^{b,k} = \min \mathcal{R}_{b,k}, \qquad t_{\max}^{b,k} = \max \mathcal{R}_{b,k}, \qquad t_{\mathrm{med}}^{b,k} = \mathrm{median}\, \mathcal{R}_{b,k},$$

and define

$$s_-^{b,k} = \max\left(t_{\mathrm{med}}^{b,k} - t_{\min}^{b,k}, 10^{-12}\right),$$
$$s_+^{b,k} = \max\left(t_{\max}^{b,k} - t_{\mathrm{med}}^{b,k}, 10^{-12}\right).$$

The reference-based alignment map is

$$A_{b,k}(t) = \begin{cases} \dfrac{1}{n_{\mathrm{cal}}} \exp\left(-\dfrac{t_{\min}^{b,k} - t}{s_-^{b,k}}\right), & t < t_{\min}^{b,k}, \\ \dfrac{1}{n_{\mathrm{cal}}} \left|\{v \in \mathcal{R}_{b,k} : v \le t\}\right|, & t_{\min}^{b,k} \le t \le t_{\max}^{b,k}, \\ 1 + \dfrac{1}{n_{\mathrm{cal}}}\left[1 - \exp\left(-\dfrac{t - t_{\max}^{b,k}}{s_+^{b,k}}\right)\right], & t > t_{\max}^{b,k}. \end{cases}$$

Within the observed normal-reference range, this is the empirical reference CDF. The reference-only tail extensions preserve raw-score ordering outside that range without using query scores to

Table 7: **Ablations of unresolved-target encoding on ADBench (57 entries).** Each entry is averaged over five random seeds before comparison. The rows form a ladder from excluding unresolved rows, to including them with a single constant target, to the graded encoding used by SCOPE. Differences and paired Wilcoxon tests are relative to the graded encoding. W/L counts entries on which the variant outperforms/underperforms the graded encoding for the corresponding metric; remaining entries are exact ties.

Unresolved rows $\mathcal{U}(N)$		AUROC			AUPRC		
in context?	target	value	Δ	W/L	value	Δ	W/L
no	—	84.28	-1.11	18/35	65.02	-2.41	14/40
yes	constant 0.5	85.01	-0.37	24/30	67.02	-0.41	24/30
yes	**graded** $[0.5, 0.75]$	**85.39**	—	—	**67.43**	—	—
p vs. graded (excluded / constant)		0.011 / 0.144			1.0×10^{-4} / 0.091		

define the transformation. In particular, the upper extension prevents all queries above the largest calibration score from collapsing to the same empirical-CDF value.

The final score for query fold $\mathcal{Q}_b$ is

$$g_b(x) = \frac{1}{K} \sum_{k=1}^{K} A_{b,k} \left(d_{b,k}(x) \right), \qquad x \in \mathcal{Q}_b.$$

F Unresolved-Target Encoding

The main method keeps unresolved target rows in the discriminative context rather than forcing them into either the normal or positive-evidence class. We further preserve their current evidence ordering by assigning graded targets in $[0.5, 0.75]$. Table 7 separates these two design choices: whether unresolved rows are included at all, and whether their relative evidence ordering is retained once they are included.

Removing unresolved rows entirely reduces the constructed task to binary $\{0, 1\}$ targets and causes a substantial decrease in performance: -1.11 AUROC and -2.41 AUPRC relative to the graded encoding. Including the same rows with a constant neutral target of 0.5 recovers most of this loss, reaching 85.01 AUROC and 67.02 AUPRC.

Preserving the relative evidence ordering of unresolved rows through graded targets provides a further 0.37 AUROC and 0.41 AUPRC improvement over the constant encoding. This additional difference is modest and is not statistically significant under the paired tests ($p = 0.144$ for AUROC and $p = 0.091$ for AUPRC). The stronger conclusion is therefore that unresolved target rows should remain represented in the context rather than being implicitly treated as normal; graded targets provide an additional way to retain their relative evidence ordering without requiring a hard label.

G Evidence and Cross-Fitting Diagnostics

G.1 Inner-fold balance

Nested cross-fitting produces refreshed scores for the two inner folds under different positive-evidence contexts. Although both folds use the same target encoding and the same normal reference, their raw score distributions need not be identical. Because the refined evidence ranking pools these scores, we examine whether this pooling systematically favors one inner fold.

Under the final SCOPE configuration, this diagnostic is available for 125 ADBench runs. For each run, we examine the composition of the deterministic top-ranked evidence set obtained from $c_b^{(R)}$. The share originating from one inner fold is 49.6% on average and 50.0% at the median, with a range of 32.4%–69.2%. In 86% of runs, this share lies within $[0.40, 0.60]$, and a test against equal contribution gives $p = 0.199$. The absolute difference between the two folds' score medians is 0.0017 on average.

Thus, we observe no systematic tendency for one inner fold to dominate the top-ranked evidence after pooling. This diagnostic concerns the deterministic ranking $c_b^{(R)}$ and is distinct from the subsequent score-weighted evidence draws $\{N_{b,k}\}_{k=1}^{K}$.

G.2 Evidence re-ranking

We additionally examine how much the discriminative refinement changes the bootstrap evidence ranking. Ground-truth anomaly labels are used only for this post-hoc diagnostic and never during evidence construction or scoring.

The initial and final top-q sets overlap by 34.2%, and the median per-entry Spearman correlation between the bootstrap and refined rankings is 0.45. Among initially selected rows that are truly anomalous, 50.7% remain selected, whereas 85.8% of initially selected rows that are actually normal are demoted. Accordingly, mean top-q precision increases from 50.4% under the bootstrap ranking to 68.8% after discriminative refinement.

These results show that refinement substantially re-orders the initial evidence rather than simply preserving the bootstrap ranking.

H Alignment and Query-Batch Effects

The alignment map $A_{b,k}$ is constructed only from $\mathcal{S}_{\mathrm{cal}}$ under the corresponding fixed context. Consequently, the alignment transformation itself does not use query scores to define its scale.

Across six diagnostic datasets, shuffling the order of queries produced exactly the same scores, with maximum absolute difference zero. The underlying frozen foundation model can nevertheless exhibit effects from changing the number of queries scored jointly.

On five of the six diagnostic datasets, scoring queries individually rather than as a full batch changed scores by at most 2×10^{-3} and yielded rank correlation $\rho \geq 0.999999$. On `mammography`, the rank correlation was 0.60 and the maximum absolute score difference was 0.33, while AUROC changed only from 89.94 to 89.99.

Thus, the reference-based alignment itself does not rely on query-cohort score normalization, but exact per-query scores can still depend on query-batch handling inside the underlying foundation model. Deployments requiring exactly reproducible scores should therefore keep the query batch configuration fixed.

I Fixed-Context Inference

The default evidence-construction procedure is transductive because unlabeled target covariates participate in bootstrap evidence construction, discriminative refinement, and final evidence resampling. Outer cross-fitting nevertheless ensures that a row in $\mathcal{Q}_b$ does not contribute to the evidence context used to score itself.

After evidence construction, the objects

$$\{\mathcal{C}_{b,k}\}_{k=1}^{K}, \qquad \{A_{b,k}\}_{k=1}^{K}$$

can be frozen and reused for later query batches without additional evidence mining, refinement, or parameter updates. The alignment maps themselves remain fixed because they depend only on the corresponding held-out normal-reference scores. As shown in Appendix H, however, exact frozen-model predictions can still depend on the query-batch configuration.

A fully inductive variant can instead construct the evidence contexts from a separate unlabeled calibration pool. In that setting, future evaluation rows are not used during bootstrap evidence construction, refinement, or evidence resampling.

J Default Hyperparameters

The default configuration used in the main experiments is

$$\begin{array}{ll} B_{\mathrm{ctx}} = 2048, & B_{\mathrm{cal}} = 256, \\ P = 3, & L = 19, \\ q = 0.05, & B_{\max} = 1024, \\ R = 3, & K = 4, \\ \tau = 0.05, & (\mu, \beta) = (0.625, 0.125), \\ k_{\mathrm{ref}}^{\min} = 2, & k_{\mathrm{fin}}^{\min} = 8. \end{array}$$

The predictive quantile levels are

$$\alpha_\ell = \frac{\ell}{20}, \qquad \ell = 1, \ldots, 19.$$

The same configuration is used across datasets unless explicitly stated otherwise.

K Reproducibility of the Comparison

This appendix documents which released benchmark records are used for comparison, how the aggregate ranking metrics are recomputed, and how irregularities in the released files are handled. Our scores, the released record files, and the aggregation scripts are included in the supplementary `repro/` directory.

K.1 Provenance of the Evaluation Sets

ADBench (57 entries). The baseline numbers are the released per-dataset records of the OutFormer evaluation, which cover 57 benchmark *entries* and 12 methods.

Our pipeline evaluates 125 dataset files because 8 benchmark entries are families of variants that the released records report as a single row: `MNIST-C` (16 variants), `MVTec-AD` (15), `CIFAR10` (10), `FashionMNIST` (10), `SVHN` (10), `20news` (6), `agnews` (4), and `amazon` (2). The remaining 52 files correspond one-to-one with benchmark entries.

Averaging within each family first gives 60 entries. Three of these—`abalone`, `arrhythmia`, and `Parkinson`—have no corresponding row in the released records and therefore cannot enter a paired comparison. They are excluded independently of our scores, leaving the 57 entries reported in the paper.

OddBench (690) and OvRBench (755). The baseline numbers are the released per-dataset records of the macrOData evaluation, which report 14 methods on 690 OddBench and 756 OvRBench rows.

For OddBench, we evaluate all 790 datasets in the distributed benchmark. One hundred have no corresponding released baseline record and are therefore excluded from the paired comparison, leaving 690 entries.

For OvRBench, the 756 released rows contain `twonorm` twice, once as `TwoNorm` and once as `twonorm`, with different reported values (e.g., OutFormer AUROC 91.00 versus 90.60). Collapsing these two rows gives 755 unique datasets, all of which we evaluate. We retain the lower-cased row; averaging the two released rows instead changes no reported aggregate to two decimal places because only one of 755 datasets is affected.

K.2 Why the Baseline Method Set Differs Between Benchmarks

The baseline sets come from two different released evaluations rather than from a method set selected by us. ADBench contributes 12 baseline methods, while OddBench and OvRBench contribute an identical set of 14. Ten methods are common to all three benchmarks.

DDPM and DeepSVDD appear only in the ADBench records, whereas OCSVM, CBLOF, EGMM, and NPT-AD appear only in the OddBench/OvRBench records. We do not re-run missing baselines, since doing so would mix the benchmark authors' released results with our own preprocessing and hyperparameter choices. Consequently, a method's average rank is comparable only within a benchmark, not directly across benchmarks.

What a dash means. A $--$ in the main comparison tables means that the method has no entry in that benchmark's released records. It does not indicate a failed run of ours. Within each source benchmark record, the baseline matrices used in the comparison are otherwise fully populated.

K.3 How the Aggregate Metrics Are Recomputed

The released files provide per-dataset performance values. Because adding SCOPE changes the comparison set, all aggregate statistics are recomputed from these per-dataset records rather than copied from published aggregate tables.

Let

$$E_\delta(m) = 1 - \frac{\mathrm{perf}_\delta(m)}{100}$$

denote the error of method m on dataset δ, and let

$$E^*_\delta = \min_m E_\delta(m)$$

denote the best error on that dataset. For n datasets and M compared methods, we use

$$\begin{aligned}
\mathrm{AvgRank}(m) &= \frac{1}{n}\sum_\delta \mathrm{rank}^{\min}(E_\delta)_m\,, \\
\mathrm{Winrate}(m) &= \frac{1}{n(M-1)}\sum_\delta \left|\{m' : E_\delta(m') > E_\delta(m)\}\right|, \\
\mathrm{rAUC}(m) &= \frac{1}{n}\sum_\delta \left(1 - \frac{E_\delta(m) - E^*_\delta}{1 - E^*_\delta}\right), \\
C_\Delta(m) &= \frac{1}{n}\sum_\delta \left(1 - \frac{E^*_\delta}{E_\delta(m)}\right).
\end{aligned}$$

For average rank, ties receive the best tied rank ($\mathrm{rank}^{\min}$), rather than the average rank. For win rate, ties contribute zero rather than one half. For rAUC, the upper error reference is 1 rather than the worst compared method. For C_Δ, the champion is defined separately for every dataset. When $E_\delta(m) = 0$, the contribution to C_Δ is defined as 0 if m is the champion and 1 otherwise.

Two preprocessing steps are applied before these statistics are computed. First, ADBench variant files belonging to one released benchmark entry are averaged to entry level before aggregation. Second, released performance values are clipped to $[0, 100]$. The ADBench records contain 5 AUROC and 4 AUPRC cells slightly above 100 (e.g., 100.01 and 100.001), while the macrOData AUPRC file contains one corrupted value (scm_structure_53, TabPFN-OD = 989.95). Clipping is required for the ratio-based aggregate metrics to remain well-defined.

K.4 Verifying the Recomputation Against Published Aggregates

We validate the definitions above on the full macrOData release: 2246 datasets $\times$ 14 methods. This is a superset of the two macrOData benchmarks reported in the main paper. Table 8 shows the largest absolute deviation over all 14 methods between our recomputation and the published aggregate table.

Seven of the eight checks agree to within 5×10^{-4}. The remaining deviation, rAUC under AUPRC, is 8.9×10^{-4}. This residual is associated with the corrupted AUPRC cell described above, whose treatment in the original aggregate computation cannot be recovered exactly. The agreement is nevertheless sufficient to recover the released aggregation definitions to the reported precision.

Elo is excluded from this check because it is order-dependent; Appendix K.6 describes our treatment of Elo.

Table 8: **Recomputation check** on the full macrOData release ($n = 2246$, $M = 14$). Entries are $\max_m |\Delta|$ between our recomputation and the published aggregate. The published table is rounded to three decimal places.

	AvgRank	Winrate	rAUC	C_Δ
AUROC	5.0×10^{-4}	4.7×10^{-4}	4.6×10^{-4}	4.4×10^{-4}
AUPRC	4.8×10^{-4}	4.8×10^{-4}	8.9×10^{-4}	4.6×10^{-4}

K.5 A Duplicated Method Column in the OvRBench Records

In the released OvRBench records, the `ICL` and `DTE-C` columns are bit-identical on all 756 released rows for both AUROC and AUPRC. This duplication is specific to the OvRBench file: the two columns differ in the OddBench and ADBench records.

For the main comparison, we retain the released method record as published rather than modifying the benchmark authors' comparison set. As a robustness check, we also recompute the OvRBench aggregates after removing one of the two identical columns. The numerical rank-based aggregates change slightly, as expected when the number of competitors changes, but the qualitative conclusions of the comparison are unchanged.

Because average rank depends directly on the number of compared methods, it should not be compared numerically across benchmarks with different method sets.

K.6 Elo: Random Orderings and Averaging

Elo is computed following the released evaluation: the initial rating is

$$R_0 = 1000,$$

the update constant is

$$K_{\mathrm{Elo}} = 32,$$

and the draw margin is

$$\varepsilon = 0.5.$$

Thus, two methods whose performance on a dataset differs by less than 0.5 percentage points receive a draw rather than a win and a loss.

All $\binom{M}{2}$ method pairs are compared on each dataset. Ratings are updated sequentially, and the dataset list is traversed 20 times with an independently sampled permutation on each pass.

Because sequential Elo updates depend on dataset order, a single ordering is not a uniquely defined aggregate statistic. We therefore repeat the complete procedure for 100 independent orderings (random seeds 0–99) and report the mean final rating. The same 100 orderings are used for every method in a comparison table.

L Computational Cost

We measure the computational cost of SCOPE on a single idle NVIDIA A100, using one process at a time with no co-tenancy. All reported wall-clock measurements use the same eight ADBench datasets, spanning 523–5250 test rows and 6–768 features. Unless otherwise stated, times are averaged across these datasets. The stage-wise measurements are instrumented within the same run and account for 99.1% of the measured wall-clock time.

Importantly, SCOPE performs no gradient-based optimization and updates no model parameter at any stage. Its computational cost consists entirely of repeated inference with the frozen tabular foundation model.

Frozen-model calls. A frozen-model call consists of context ingestion (`fit` in our implementation) followed by query scoring (`predict`). Here, `fit` does not denote parameter fitting: it only supplies the in-context examples to the frozen predictor, with no gradient computation or parameter

Table 9: **Stage-wise computational cost.** Wall-clock time in seconds, averaged over eight ADBench datasets on a single idle NVIDIA A100. All stages use frozen-model inference only.

Configuration	T_{boot}	T_{refine}	T_{final}	Total
Bootstrap s_0 only	13.13	—	—	13.13
R=3, K=1	13.61	14.28	5.96	34.16
Discriminative (R=0, K=4)	13.04	0.00	23.80	37.12
R=1, K=4	13.13	4.81	23.82	42.04
Full SCOPE (R=3, K=4)	13.13	14.25	23.85	51.53
R=5, K=4	13.27	24.14	24.10	61.81

Table 10: **Context construction and subsequent scoring.** Times are averaged over the same eight datasets. The construction row includes context ingestion and all intermediate scoring required to obtain the final evidence contexts.

Computation	Time (s)
Full SCOPE procedure	51.53
Context construction and intermediate scoring	50.01
Final batched scoring with fixed contexts	1.52

update. Across the eight datasets, context ingestion takes $1.244\,\text{s}$ and query scoring $0.379\,\text{s}$ per call on average, so most of the cost of a call arises from processing the context rather than scoring the query batch.

For fixed P, R, and K, the number of frozen-model calls is determined by

$$N_{\text{call}} = 2P + 4R + 4K.$$

The default configuration $P = 3$, $R = 3$, and $K = 4$ therefore uses 34 calls. Because the number of probed features P is fixed rather than proportional to the number of columns, this call count does not increase with feature dimensionality.

Stage-wise cost. Table 9 decomposes the total wall-clock time into bootstrap evidence construction, cross-fitted refinement, and final evidence-context scoring. The measurements follow the structure of the algorithm: bootstrap time is nearly unchanged across configurations, refinement cost increases with R, and final-context cost increases with K.

Under the default configuration, the final evidence-context stage is the largest component of measured wall-clock time, followed by refinement and bootstrap. Refinement adds approximately $4.8\,\text{s}$ per round in this measurement, while reducing the number of final evidence contexts from $K = 4$ to $K = 1$ decreases the final-stage time from approximately $24\,\text{s}$ to $6\,\text{s}$. Thus, both R and K provide direct controls over computational cost.

Context construction and subsequent scoring. The full default procedure takes $51.53\,\text{s}$ per dataset on average. Most of this time is spent constructing and refining the evidence contexts. Once the final evidence contexts and their alignment maps have been constructed, subsequent batched prediction requires only the $K = 4$ final scoring calls, which take $1.52\,\text{s}$ in total on average across the measured datasets.

The cost of the final scoring stage is primarily incurred per query batch rather than per individual query. Across the measured datasets, batched scoring corresponds to 0.16–3.3 ms per query (median 2.7 ms). This favors batched evaluation when multiple observations are available for scoring.

Scaling behavior. Absolute runtime varies across datasets because the cost of each frozen-model call depends on the size and dimensionality of the tabular input. Table 11 summarizes the observed stage costs after grouping the eight timing datasets by dataset size.

Across these measurements, absolute runtime tends to increase with feature dimensionality. For example, `23_mammography` has 5114 rows but only $d = 6$ features and takes $16.7\,\text{s}$, whereas `25_musk` has fewer rows (1580) but $d = 166$ features and takes $40.1\,\text{s}$. Fixing $P = 3$ prevents

Table 11: **Scaling of full SCOPE.** Wall-clock times in seconds on the same eight ADBench datasets, grouped by dataset size.

Size	# datasets	med. d	med. n_{test}	T_{boot}	T_{refine}	T_{final}	Total
Small	3	9	525	3.43	4.55	6.28	14.4
Medium	3	166	1874	11.28	13.11	21.94	46.6
Large	2	387	5182	30.44	30.51	53.07	114.6

Table 12: **Effect of verified positive evidence on the 53 OddBench datasets with anomaly prevalence below 1%.** Differences are relative to the fully label-free SCOPE configuration ($k = 0$). W/L reports dataset-level wins/losses relative to $k = 0$; remaining datasets are ties.

k	AUROC	AUPRC	ΔAUROC	ΔAUPRC	W/L
0	73.31	18.01	—	—	—
1	75.23	20.77	+1.91	+2.76	38/12
2	76.66	23.77	+3.35	+5.76	42/7
3	76.53	25.12	+3.21	+7.11	38/11
5	78.13	26.21	+4.82	+8.20	46/4
7	78.63	27.31	+5.32	+9.30	47/3
10	79.53	27.49	+6.22	+9.48	44/6
all	80.13	28.30	+6.81	+10.29	44/6

the *number* of bootstrap calls from growing with dimensionality; increasing dimensionality instead affects the cost of each individual frozen-model call.

M Verified Positive Feedback under Extreme Anomaly Rarity

The main experiments assume that no labeled anomalies are available. Here we study a complementary setting in which a small number of anomalies become verified after the initial label-free detection step. Because SCOPE represents anomalous observations through in-context positive evidence rather than task-specific model parameters, these verified examples can be incorporated directly into the final positive-evidence contexts without retraining the frozen backbone.

We focus on the 53 OddBench datasets with contamination below 1%. Two partially matched entries, `RestaurantInspections` and `WebTrafficError`, are excluded from this analysis. Across the resulting datasets, anomaly prevalence ranges from 0.09% to 0.97% (median 0.46%), and the evidence-mining fold contains between 2 and 29 true anomalies (median 7). For each value of k, we add k verified anomalies to the final positive-evidence context while retaining the self-mined evidence used by the original method.

Effect of a small verified-positive budget. Table 12 shows that even a small number of verified positives leads to consistent improvements over the fully label-free configuration ($k = 0$).

With only one verified positive, AUROC increases by 1.91 points and AUPRC by 2.76 points, with 38 wins and 12 losses across the 53 datasets ($p = 3.2 \times 10^{-5}$ for AUROC and $p = 4.0 \times 10^{-3}$ for AUPRC). Two verified positives increase the gains to +3.35 AUROC and +5.76 AUPRC ($p = 5.5 \times 10^{-7}$ and 1.4×10^{-6}, respectively). The improvement continues as additional verified positives become available, although the trajectory is not strictly monotonic at every value of k.

Comparison with published baselines. The effect is also visible in the aggregate ranking among the 15 methods available on these datasets. Without verified positives, SCOPE has an AUROC average rank of 7.04, compared with 6.58 for OutFormer. A single verified positive improves the SCOPE rank to 6.27, while two improve it further to 5.64. Thus, one verified anomaly is sufficient for SCOPE to obtain a better AUROC average rank than the strongest published baseline in this low-prevalence subset.

Table 13: **Average-rank progression under verified positive feedback.** Lower is better. The reference baseline is the strongest published method for the corresponding metric in this subset.

	$k=0$	$k=1$	$k=2$	$k=5$	$k=7$	$k=\text{all}$
AUROC: SCOPE	7.04	6.27	5.64	5.53	5.40	5.09
AUROC: OutFormer	6.58	6.64	6.71	6.73	6.73	6.73
AUPRC: SCOPE	7.62	—	—	5.78	5.33	4.76
AUPRC: OCSVM	5.60	—	—	5.73	5.76	5.89

Table 14: **AUROC gain as a function of the number of true anomalies available in the evidence-mining fold.** Values are differences relative to the fully label-free configuration.

k	≤ 4 (14)	4–8 (19)	8–15 (13)	>15 (7)
1	+2.10	+2.08	+2.22	+0.53
2	+5.03	+2.65	+3.73	+1.18
5	+5.42	+3.74	+6.08	+4.22
10	+7.68	+2.59	+9.97	+6.14
all	+7.68	+2.58	+10.32	+10.03

For AUPRC, more verified evidence is required. The label-free configuration has an average rank of 7.62, whereas the strongest published baseline, OCSVM, has rank 5.60. The gap narrows as verified positives are added, and SCOPE reaches rank 5.33 at $k = 7$, compared with 5.76 for OCSVM.

Dependence on evidence scarcity. We next group datasets according to the number of true anomalies available in the evidence-mining fold. Table 14 reports the AUROC gain over $k = 0$.

The largest early gains appear when very few true anomalies are available: among the 14 datasets containing at most four true anomalies in the mining fold, two verified positives improve AUROC by 5.03 points on average. The relationship is not strictly monotonic across all groups and values of k, but the result shows that a very small verified-positive budget can provide substantial benefit precisely when self-constructed positive evidence is scarce.

Oracle positive evidence. Finally, the $k = \text{all}$ condition provides an oracle diagnostic in which all available true anomalies are supplied as positive evidence. Relative to the label-free configuration, this increases average AUROC by 6.81 points and AUPRC by 10.29 points. This substantial gap indicates that positive-evidence identification remains an important source of headroom in the extreme-rarity regime.

The oracle does not, however, resolve every dataset. Some tasks remain difficult even when the positive evidence is known, indicating that errors can arise not only from evidence construction but also from the subsequent discriminative in-context prediction. Conversely, on some datasets the label-free procedure is already close to the oracle, indicating that self-constructed evidence can be sufficient without verified feedback. Taken together, these results separate two sources of error—identifying useful positive evidence and discriminating from that evidence—while showing that SCOPE can directly exploit verified positives when they become available without updating the underlying model.

Table 15: Extended Results (AUPRC). Baseline entries retain the reported mean±std; SCOPE uses the final evaluation mean, with the standard deviation.

Dataset	kNN	LOF	IForest	ICL	DeepSVDD	GOAD	DDPM	DTE-NP	DTE-C	FoMo-0D	TabPFN-OD	OutFormer	SCOPE
hepatitis	0.9031±0.0436	0.4367±0.1079	0.5536±0.0609	0.9983±0.0038	0.9873±0.0105	0.6577±0.0543	0.9514±0.0134	0.8232±0.1026	0.9582±0.0385	0.9945±0.0110	0.9988±0.0023	0.9957±0.0086	**1.0000**±0.0000
MVTec-AD	0.7576±0.0271	0.7579±0.0295	0.7000±0.0307	0.8946±0.0223	0.8378±0.0332	0.7262±0.0274	0.7366±0.0272	0.8294±0.0268	0.8511±0.0296	0.7325±0.0179	0.8891±0.0102	0.8378±0.0054	**0.9072**±0.0044
wine	0.9511±0.0181	0.8995±0.0264	0.6712±0.0762	0.9826±0.0389	0.7856±0.0959	0.7010±0.0629	0.9765±0.0072	0.9680±0.0558	0.9985±0.0023	0.9962±0.0077	0.9987±0.0027	0.9971±0.0058	**1.0000**±0.0000
lymphography	0.9917±0.0094	0.8416±0.0427	0.9438±0.0328	**1.0000**±0.0000	0.9682±0.0369	0.9876±0.0088	0.9930±0.0102	0.9934±0.0093	0.8677±0.0924	0.9915±0.0126	0.9935±0.0130	0.9976±0.0049	0.9878±0.0000
wpbc	0.4611±0.0274	0.4120±0.0262	0.4073±0.0315	0.8931±0.0537	0.7488±0.0558	0.3888±0.0382	0.5461±0.0375	0.6902±0.1391	0.6035±0.0488	0.6736±0.0434	0.8920±0.0496	0.8134±0.0406	**0.9118**±0.0055
glass	0.4232±0.0847	0.3812±0.0991	0.2137±0.0358	0.9235±0.0832	0.5235±0.2104	0.1833±0.0723	0.3121±0.1130	0.3738±0.1551	0.4151±0.0591	0.8360±0.0463	0.8413±0.0757	0.7176±0.1113	**0.9867**±0.0004
wbc	0.9201±0.0396	0.2489±0.0349	0.9424±0.0395	0.9512±0.0514	0.5651±0.1112	0.9195±0.0326	0.9384±0.0245	0.9610±0.0217	0.2997±0.0310	0.9646±0.0249	0.9698±0.0271	0.9487±0.0544	**0.9973**±0.0067
vertebral	0.2611±0.0249	0.3387±0.0430	0.2075±0.0198	0.5875±0.0730	0.2342±0.0317	0.2139±0.0498	0.3584±0.0927	0.2521±0.0890	0.3514±0.0539	0.6939±0.0397	0.7713±0.0835	0.6628±0.0389	**0.8518**±0.0103
stamps	0.7168±0.0835	0.6484±0.0817	0.5884±0.0684	0.7954±0.0544	0.4262±0.0994	0.4957±0.1772	0.6474±0.1287	0.8247±0.0408	0.5765±0.0840	0.8937±0.0350	0.8460±0.0454	0.9095±0.0510	**0.9806**±0.0034
ionosphere	0.9795±0.0069	0.9458±0.0157	0.9170±0.0189	0.9906±0.0032	0.9809±0.0081	0.9317±0.0264	0.9643±0.0050	0.9822±0.0102	0.9683±0.0038	0.9775±0.0076	0.9915±0.0022	0.9846±0.0043	**0.9936**±0.0003
wdbc	0.8203±0.0328	0.9364±0.0306	0.7198±0.0857	0.9560±0.0612	0.8432±0.0889	0.7877±0.0426	0.8430±0.0444	0.9047±0.0791	0.6882±0.1242	0.9382±0.0830	0.8501±0.1018	0.9040±0.0816	**1.0000**±0.0000
20news	0.1347±0.0052	0.1504±0.0067	0.1156±0.0043	0.1462±0.0088	0.1290±0.0190	0.1149±0.0051	0.1148±0.0044	0.1561±0.0135	0.1733±0.0236	0.1351±0.0084	**0.2096**±0.0069	0.1886±0.0148	0.1788±0.0192
breastw	0.9892±0.0032	0.8001±0.1009	0.9949±0.0012	0.9679±0.0161	0.9601±0.0128	0.9877±0.0037	0.9860±0.0051	0.9919±0.0014	0.8825±0.0106	0.9903±0.0034	0.9946±0.0020	0.9923±0.0023	**0.9954**±0.0009
pima	0.7537±0.0299	0.6840±0.0379	0.7365±0.0207	0.7863±0.0194	0.5975±0.0175	0.6515±0.0880	0.7118±0.0206	0.7970±0.0244	0.6798±0.0286	0.7927±0.0174	0.8054±0.0167	0.8005±0.0302	**0.8370**±0.0053
vowels	0.3021±0.0000	0.3309±0.0000	0.1197±0.0105	0.2739±0.0575	0.1688±0.0193	0.2094±0.0243	0.4272±0.0480	0.3159±0.0159	0.3810±0.0489	0.2410±0.0000	0.1906±0.0089	**0.4678**±0.0000	0.0883±0.0050
yeast	0.4826±0.0000	0.4894±0.0000	0.4678±0.0037	0.4955±0.0136	0.4921±0.0388	0.5077±0.0218	0.5105±0.0165	0.4812±0.0049	0.4974±0.0072	0.5100±0.0000	0.4973±0.0008	0.5210±0.0000	**0.6935**±0.0114
letter	0.0870±0.0000	0.1126±0.0000	0.0822±0.0025	**0.1280**±0.0117	0.0893±0.0052	0.0813±0.0005	0.0953±0.0051	0.0857±0.0013	0.0895±0.0013	0.0933±0.0000	0.0944±0.0002	0.0945±0.0000	0.0877±0.0004
cardio	0.7722±0.0000	0.7015±0.0000	0.7863±0.0269	0.4791±0.1147	0.3889±0.0552	**0.8479**±0.0057	0.6928±0.0091	0.7741±0.0085	0.6929±0.0127	0.7892±0.0000	0.5514±0.0188	0.7185±0.0000	0.5382±0.0103
fault	0.6198±0.0000	0.5044±0.0000	0.5919±0.0202	0.6318±0.0070	0.5546±0.0148	0.6214±0.0082	**0.6475**±0.0069	0.6217±0.0014	0.6393±0.0072	0.6308±0.0000	0.6293±0.0036	0.6164±0.0000	0.5559±0.0178
internetads	0.4922±0.0000	0.5043±0.0000	0.2920±0.0174	0.6003±0.0139	0.5156±0.0475	0.4743±0.0086	0.4770±0.0016	0.5130±0.0202	0.5522±0.0365	0.3829±0.0522	0.4590±0.0380	0.6952±0.0170	**0.8270**±0.0043
cardiotocography	0.5743±0.0000	0.5732±0.0000	0.6285±0.0342	0.4866±0.0384	0.4578±0.0510	**0.6752**±0.0082	0.5131±0.0198	0.5868±0.0114	0.5334±0.0126	0.6355±0.0000	0.4704±0.0033	0.5689±0.0000	0.4606±0.0070
musk	1.0000±0.0000	1.0000±0.0000	0.4039±0.2610	0.9221±0.0632	0.9991±0.0017	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	0.9700±0.0227	1.0000±0.0000	1.0000±0.0000	0.5998±0.0766
waveform	0.2700±0.0000	0.3066±0.0000	0.1053±0.0076	0.1863±0.0408	0.1152±0.0339	0.0886±0.0064	0.0931±0.0105	0.2787±0.0000	0.0999±0.0113	0.0995±0.0000	0.2082±0.0039	0.1365±0.0005	**0.7161**±0.0565
speech	0.0280±0.0000	0.0315±0.0000	0.0325±0.0100	0.0338±0.0050	0.0338±0.0038	0.0281±0.0031	0.0300±0.0029	0.0317±0.0000	0.0285±0.0012	0.0294±0.0031	0.0281±0.0032	0.0275±0.0003	**0.0356**±0.0005
thyroid	0.8094±0.0000	0.6057±0.0000	0.7966±0.0562	0.5151±0.1275	0.6906±0.0810	0.8009±0.0089	0.8222±0.0087	0.8103±0.0031	0.8167±0.0097	0.6700±0.0000	0.8160±0.0014	0.7013±0.0033	**0.9422**±0.0039
spambase	0.8332±0.0000	0.7271±0.0000	0.8826±0.0132	0.8678±0.0059	0.7526±0.0244	0.8209±0.0021	0.7289±0.0042	0.8365±0.0058	0.8380±0.0052	0.8099±0.0000	0.8592±0.0006	0.7968±0.0007	**0.9616**±0.0020
wilt	0.1225±0.0000	0.1574±0.0000	0.0881±0.0051	0.2894±0.0331	0.0708±0.0017	0.1086±0.0137	0.1724±0.0046	0.1220±0.0160	0.2541±0.0136	**0.7746**±0.0000	0.2403±0.0009	0.5606±0.0053	0.7726±0.0161
SVHN	0.1534±0.0000	0.1597±0.0000	0.1385±0.0049	0.1560±0.0031	0.1240±0.0095	0.1493±0.0018	0.1508±0.0004	0.1553±0.0000	0.1547±0.0047	0.1467±0.0033	0.1567±0.0050	0.1671±0.0007	**0.2175**±0.0230
optdigits	0.2911±0.0000	0.4363±0.0000	0.1541±0.0321	0.5094±0.0859	0.0453±0.0104	0.0776±0.0115	0.2556±0.0425	0.3175±0.0486	0.1534±0.0217	0.3188±0.0000	0.2843±0.0180	0.1737±0.0009	**0.9900**±0.0001
CIFAR10	0.1962±0.0000	0.2217±0.0000	0.1646±0.0064	0.1739±0.0059	0.1403±0.0108	0.1940±0.0040	0.1955±0.0008	0.1991±0.0000	0.1968±0.0086	0.1745±0.0052	0.1858±0.0055	0.2188±0.0011	**0.2683**±0.0311
pageblocks	0.6760±0.0000	**0.7107**±0.0000	0.4342±0.0200	0.6811±0.0230	0.5205±0.0389	0.6350±0.0117	0.6210±0.0095	0.6745±0.0010	0.6642±0.0123	0.6267±0.0000	0.7010±0.0062	0.6381±0.0043	0.5015±0.0187
satimage-2	**0.9669**±0.0000	0.8846±0.0000	0.9453±0.0055	0.9470±0.0119	0.7628±0.0821	0.9589±0.0010	0.8805±0.0510	0.9616±0.0000	0.6821±0.0315	0.9278±0.0000	0.5524±0.0154	0.9218±0.0017	0.2703±0.0703
FashionMNIST	0.5915±0.0000	0.6361±0.0000	0.4473±0.0188	0.6308±0.0096	0.4510±0.0204	0.5659±0.0040	0.5714±0.0012	0.5979±0.0000	0.5501±0.0104	0.5219±0.0278	0.5841±0.0151	0.6417±0.0031	**0.8453**±0.0042
landsat	0.5485±0.0000	0.6137±0.0000	0.4731±0.0350	0.5312±0.0157	0.4943±0.0240	0.3121±0.0080	0.3483±0.0091	0.5452±0.0405	0.3675±0.0123	0.5824±0.0000	0.4218±0.0013	0.3862±0.0007	**0.6503**±0.0074
satellite	0.8601±0.0000	0.8586±0.0000	0.8235±0.0088	0.8762±0.0024	0.8110±0.0197	0.7896±0.0046	0.8509±0.0018	0.8583±0.0072	0.8479±0.0035	**0.8805**±0.0000	0.8372±0.0005	0.8041±0.0005	0.8713±0.0016
pendigits	0.9699±0.0000	0.7855±0.0000	0.5879±0.0521	0.6641±0.0758	0.0934±0.0778	0.3335±0.0285	0.6114±0.0473	0.9189±0.0330	0.4844±0.0592	0.6633±0.0000	0.6158±0.0087	0.8307±0.0064	**0.9984**±0.0006
annthyroid	0.6807±0.0000	0.5353±0.0000	0.5902±0.0539	0.4583±0.0216	0.2783±0.0593	0.5874±0.0500	0.6288±0.0259	0.6815±0.0038	**0.8288**±0.0059	0.4881±0.0000	0.6843±0.0008	0.6912±0.0034	0.5583±0.0610
mnist	0.7272±0.0000	0.7097±0.0000	0.5415±0.0653	0.6845±0.0163	0.4600±0.0955	0.6509±0.0057	0.6242±0.0439	0.7368±0.0131	0.5626±0.0326	0.5797±0.0000	0.7690±0.0142	0.6123±0.0034	**0.9721**±0.0013
MNIST-C	0.4620±0.0000	0.5189±0.0000	0.3280±0.0243	0.5147±0.0110	0.3144±0.0304	0.4123±0.0035	0.4178±0.0012	0.4742±0.0000	0.4716±0.0136	0.3840±0.0334	0.5126±0.0133	0.5219±0.0047	**0.8085**±0.0106
agnews	0.1668±0.0000	0.2586±0.0000	0.1194±0.0032	0.1542±0.0038	0.1022±0.0185	0.1242±0.0031	0.1185±0.0003	0.1735±0.0000	0.1922±0.0297	0.1516±0.0317	0.2397±0.0089	0.2050±0.0020	**0.4094**±0.0187
amazon	0.1169±0.0000	0.1104±0.0000	0.1107±0.0019	0.1019±0.0008	0.1024±0.0127	0.1094±0.0021	0.1076±0.0002	0.1173±0.0000	0.1115±0.0065	0.1110±0.0146	0.1206±0.0013	0.1148±0.0011	**0.1380**±0.0025
imdb	0.0892±0.0000	0.0903±0.0000	0.0897±0.0017	0.1024±0.0013	0.0968±0.0147	0.0880±0.0011	0.0870±0.0002	0.0897±0.0000	0.0890±0.0052	0.1053±0.0236	0.0925±0.0032	0.0903±0.0017	**0.1076**±0.0063
yelp	0.1603±0.0000	0.1614±0.0000	0.1315±0.0029	0.1040±0.0009	0.1002±0.0101	0.1313±0.0039	0.1250±0.0002	0.1635±0.0000	0.1300±0.0157	0.1390±0.0300	0.1666±0.0020	0.1581±0.0022	**0.1705**±0.0284
mammography	0.4127±0.0000	0.3407±0.0000	0.3794±0.0321	0.1711±0.0375	0.2754±0.1140	0.2782±0.0384	0.1993±0.0092	0.4209±0.0086	0.3980±0.0426	0.3628±0.0126	0.4743±0.0032	0.3786±0.0063	**0.7154**±0.0151
magic.gamma	0.8586±0.0000	0.8636±0.0000	0.8027±0.0107	0.8133±0.0057	0.6954±0.0073	0.7613±0.0242	0.8797±0.0084	0.8615±0.0074	0.8968±0.0050	0.8754±0.0018	0.8854±0.0005	0.9007±0.0007	**0.9476**±0.0004
campaign	0.4904±0.0000	0.4024±0.0000	0.4573±0.0185	0.4890±0.0098	0.3695±0.1270	0.2309±0.0718	0.4887±0.0029	0.4995±0.0067	0.4690±0.0071	0.3424±0.0047	0.4323±0.0090	0.4767±0.0017	**0.7155**±0.0039
shuttle	0.9786±0.0000	0.9975±0.0000	0.9861±0.0034	0.9972±0.0014	0.9803±0.0013	0.6016±0.2692	0.9791±0.0026	0.9814±0.0048	0.9403±0.0011	0.9936±0.0017	0.9834±0.0006	0.9806±0.0033	**0.9981**±0.0023
aloi	0.0602±0.0000	0.0654±0.0000	0.0582±0.0010	0.0550±0.0013	0.0623±0.0035	0.0570±0.0024	0.0597±0.0006	0.0605±0.0006	0.0576±0.0004	0.0693±0.0005	0.0589±0.0002	0.0635±0.0002	**0.0926**±0.0040
smtp	**0.5053**±0.0592	0.4809±0.0738	0.0110±0.0012	0.0381±0.0383	0.3073±0.2282	0.3240±0.0848	0.4081±0.1336	0.5020±0.0639	0.5037±0.0615	0.3818±0.0990	0.3683±0.0484	0.5001±0.0579	0.0144±0.0187
backdoor	0.4654±0.0141	0.5347±0.0260	0.0937±0.0148	0.8918±0.0104	0.8477±0.0279	0.0631±0.0194	0.1420±0.0063	0.4570±0.1250	0.6244±0.0239	0.4386±0.1497	0.2084±0.0398	0.7197±0.0132	**0.9342**±0.0038
celeba	0.1192±0.0050	0.0361±0.0015	0.1170±0.0135	0.0974±0.0068	0.0709±0.0432	0.0401±0.0121	0.1803±0.0260	0.1065±0.0049	0.1419±0.0231	0.0839±0.0038	0.1380±0.0073	0.0664±0.0026	**0.2039**±0.0125
skin	0.9824±0.0041	0.6168±0.0185	0.6458±0.0109	0.3246±0.0100	0.4299±0.0328	0.4218±0.0184	0.7637±0.0579	0.9478±0.0233	0.6908±0.0050	0.5104±0.0172	0.6880±0.0913	0.8831±0.0043	**0.9965**±0.0001
fraud	0.3868±0.0719	0.5509±0.0819	0.1822±0.0366	0.5388±0.0878	0.4833±0.1713	0.2944±0.2466	0.6921±0.0546	0.4210±0.0763	0.6214±0.1092	0.5513±0.1377	0.4463±0.0773	0.2252±0.0233	**0.7403**±0.0071
cover	0.5579±0.0374	0.8292±0.0219	0.0866±0.0153	0.3448±0.1639	0.0269±0.0153	0.0109±0.0018	0.7327±0.0336	0.5997±0.1057	0.6373±0.1233	0.7257±0.0838	0.0269±0.0064	0.8415±0.0300	**0.9055**±0.0080
census	0.2168±0.0063	0.1371±0.0042	0.1419±0.0073	0.2120±0.0061	0.1535±0.0107	0.0869±0.0099	0.1967±0.0060	0.2105±0.0067	0.1794±0.0109	0.1601±0.0297	0.1752±0.0253	0.1840±0.0121	**0.4478**±0.0221
http	1.0000±0.0000	0.9712±0.0392	0.5343±0.1208	0.7082±0.4206	0.3609±0.3185	0.6838±0.0801	0.9996±0.0006	0.9710±0.0404	0.5545±0.0561	0.9058±0.0501	0.8048±0.0738	0.9893±0.0081	1.0000±0.0000
donors	0.8909±0.0094	0.6339±0.0189	0.4051±0.0359	0.9835±0.0087	0.4275±0.2748	0.0900±0.0193	0.2666±0.0291	0.8555±0.0456	0.7133±0.0389	0.9843±0.0048	0.5582±0.0508	0.8112±0.0095	**0.9978**±0.0002
rAUC	0.754	0.715	0.600	0.740	0.606	0.602	0.712	0.768	0.714	0.758	0.744	0.794	**0.915**
WinRate	0.538	0.493	0.266	0.569	0.249	0.231	0.431	0.607	0.472	0.529	0.602	0.633	**0.829**
Avg. Rank	6.40	6.95	9.79	6.16	9.96	10.09	7.70	5.58	7.21	6.65	5.61	5.25	**3.04**

Table 16: Extended Results (AUROC). Baseline entries retain the reported mean±std; SCOPE uses the final evaluation mean, with the standard deviation.

Dataset	kNN	LOF	IForest	ICL	DeepSVDD	GOAD	DDPM	DTE-NP	DTE-C	FoMo-0D	TabPFN-OD	OutFormer	SCOPE
hepatitis	0.9646±0.0146	0.6692±0.0701	0.8269±0.0275	0.9994±0.0013	0.9957±0.0024	0.8450±0.0325	0.9774±0.0118	0.9322±0.0390	0.9878±0.0088	0.9988±0.0024	0.9996±0.0009	0.9989±0.0022	**1.0000**±0.0000
MVTec-AD	0.8151±0.0182	0.8036±0.0214	0.7738±0.0195	**0.9475**±0.0084	0.8955±0.0219	0.7706±0.0210	0.7800±0.0199	0.8966±0.0151	0.8939±0.0200	0.7531±0.0126	0.9422±0.0046	0.8630±0.0029	0.9432±0.0024
wine	0.9919±0.0022	0.9836±0.0038	0.9392±0.0179	0.9987±0.0029	0.9216±0.0434	0.9411±0.0186	0.9961±0.0009	0.9944±0.0097	0.9997±0.0004	0.9994±0.0012	0.9997±0.0005	0.9995±0.0010	**1.0000**±0.0000
lymphography	0.9993±0.0008	0.9821±0.0075	0.9945±0.0032	**1.0000**±0.0000	0.9973±0.0030	0.9989±0.0008	0.9994±0.0009	0.9993±0.0009	0.9899±0.0040	0.9993±0.0011	0.9995±0.0011	0.9998±0.0004	0.9990±0.0000
wpbc	0.6367±0.0234	0.5736±0.0201	0.5633±0.0272	**0.9661**±0.0117	0.8267±0.0547	0.5139±0.0574	0.6649±0.0317	0.8316±0.1346	0.6888±0.0270	0.7610±0.0211	0.9650±0.0153	0.9073±0.0161	0.9613±0.0019
glass	0.9204±0.0112	0.8882±0.0198	0.8109±0.0256	0.9944±0.0058	0.8367±0.1620	0.5903±0.1264	0.6667±0.1363	0.8964±0.0354	0.9242±0.0230	0.9830±0.0053	0.9898±0.0057	0.9450±0.0289	**0.9989**±0.0000
wbc	0.9912±0.0019	0.8052±0.0447	0.9941±0.0037	0.9966±0.0036	0.9144±0.0364	0.9914±0.0020	0.9920±0.0039	0.9954±0.0028	0.8053±0.0715	0.9961±0.0031	0.9965±0.0037	0.9955±0.0050	**0.9998**±0.0010
vertebral	0.5767±0.0358	0.6430±0.0131	0.4564±0.0383	0.7919±0.0506	0.4479±0.0171	0.4673±0.0874	0.7067±0.0628	0.5430±0.1547	0.6641±0.0190	0.8720±0.0258	0.9530±0.0182	0.8477±0.0256	**0.9598**±0.0027
stamps	0.9589±0.0144	0.9374±0.0236	0.9347±0.0142	0.9668±0.0109	0.7109±0.0368	0.8146±0.1536	0.9184±0.0415	0.9787±0.0037	0.9160±0.0204	0.9752±0.0039	0.9773±0.0059	0.9859±0.0057	**0.9964**±0.0007
ionosphere	0.9744±0.0098	0.9429±0.0220	0.9121±0.0137	0.9898±0.0032	0.9720±0.0126	0.9154±0.0305	0.9460±0.0085	0.9777±0.0139	0.9542±0.0058	0.9659±0.0125	0.9904±0.0024	0.9795±0.0086	**0.9920**±0.0005
wdbc	0.9905±0.0027	0.9962±0.0021	0.9873±0.0064	0.9978±0.0027	0.9931±0.0040	0.9896±0.0025	0.9930±0.0017	0.9952±0.0038	0.9848±0.0077	0.9977±0.0025	0.9935±0.0037	0.9964±0.0026	**1.0000**±0.0000
20news	0.5736±0.0066	0.6019±0.0062	0.5491±0.0117	0.6130±0.0105	0.5564±0.0417	0.5492±0.0104	0.5488±0.0052	0.5997±0.0074	0.6430±0.0317	0.5693±0.0146	**0.6738**±0.0039	0.6543±0.0081	0.6533±0.0140
breastw	0.9905±0.0026	0.8891±0.0680	0.9950±0.0008	0.9828±0.0045	0.9696±0.0092	0.9886±0.0033	0.9870±0.0043	0.9928±0.0012	0.9278±0.0175	0.9915±0.0023	0.9951±0.0013	0.9939±0.0012	**0.9960**±0.0005
pima	0.7694±0.0187	0.7053±0.0219	0.7426±0.0163	0.7968±0.0306	0.5799±0.0285	0.6231±0.1370	0.7027±0.0228	0.8150±0.0257	0.6988±0.0195	0.8046±0.0148	0.8281±0.0111	0.8130±0.0115	**0.8672**±0.0036
vowels	0.8221±0.0000	0.8630±0.0000	0.6183±0.0662	0.8510±0.0213	0.5573±0.0443	0.6849±0.0371	0.8638±0.0194	0.8142±0.0144	0.8693±0.0225	0.8129±0.0000	0.7610±0.0058	**0.8714**±0.0000	0.6406±0.0159
yeast	0.4474±0.0000	0.4579±0.0000	0.4180±0.0075	0.4898±0.0239	0.4762±0.0596	0.5253±0.0388	0.4913±0.0285	0.4458±0.0033	0.4708±0.0110	0.4991±0.0000	0.4892±0.0011	0.5115±0.0000	**0.6928**±0.0090
letter	0.3543±0.0000	**0.4483**±0.0000	0.3204±0.0164	0.4268±0.0117	0.3640±0.0305	0.3108±0.0055	0.3805±0.0112	0.3438±0.0098	0.3672±0.0095	0.4090±0.0000	0.4217±0.0017	0.3833±0.0000	0.3755±0.0052
cardio	0.9200±0.0000	0.9221±0.0000	0.9332±0.0143	0.8001±0.0212	0.6543±0.0437	**0.9601**±0.0027	0.8694±0.0196	0.9180±0.0059	0.8726±0.0100	0.9427±0.0000	0.7626±0.0034	0.9126±0.0000	0.7652±0.0085
fault	0.5873±0.0000	0.4742±0.0000	0.5586±0.0203	0.6063±0.0050	0.5431±0.0162	0.5889±0.0061	0.6109±0.0116	0.5864±0.0065	0.5946±0.0145	**0.6182**±0.0000	0.5668±0.0052	0.5836±0.0000	0.4742±0.0156
internetads	0.6808±0.0000	0.7172±0.0000	0.4787±0.0211	0.7220±0.0055	0.7296±0.0324	0.6565±0.0021	0.6576±0.0006	0.6996±0.0222	0.7757±0.0154	0.5564±0.0712	0.6069±0.0122	0.7698±0.0114	**0.8576**±0.0023
cardiotocography	0.6211±0.0000	0.6449±0.0000	0.7424±0.0288	0.5420±0.0180	0.4775±0.0859	**0.7606**±0.0140	0.5454±0.0296	0.6376±0.0188	0.6013±0.0259	0.7215±0.0000	0.4738±0.0056	0.6229±0.0000	0.4995±0.0074
musk	1.0000±0.0000	1.0000±0.0000	0.9058±0.0620	0.9937±0.0063	0.9999±0.0001	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	1.0000±0.0000	0.9962±0.0043	1.0000±0.0000	1.0000±0.0000	0.7899±0.0081
waveform	0.7521±0.0000	0.7600±0.0000	0.7229±0.0152	0.6868±0.0366	0.5994±0.0258	0.6499±0.0310	0.6217±0.0264	0.7448±0.0000	0.6521±0.0105	0.7102±0.0000	0.7152±0.0010	0.7263±0.0007	**0.9456**±0.0078
speech	0.3636±0.0000	0.3753±0.0000	0.3770±0.0172	0.4886±0.0272	0.4888±0.0288	0.3663±0.0114	0.3696±0.0086	0.4137±0.0000	0.3817±0.0057	0.4010±0.0175	0.3718±0.0122	0.3739±0.0024	**0.5007**±0.0053
thyroid	0.9868±0.0000	0.9272±0.0000	0.9896±0.0020	0.9540±0.0098	0.8877±0.0369	0.9515±0.0085	0.9795±0.0022	0.9863±0.0004	0.9874±0.0014	0.9684±0.0000	0.9761±0.0006	0.9734±0.0005	**0.9942**±0.0009
spambase	0.8336±0.0000	0.7323±0.0000	0.8518±0.0169	0.8353±0.0045	0.7024±0.0502	0.8178±0.0036	0.6454±0.0080	0.8374±0.0069	0.8301±0.0041	0.7821±0.0000	0.8554±0.0008	0.7973±0.0009	**0.9465**±0.0018
wilt	0.6366±0.0000	0.6881±0.0000	0.4797±0.0312	0.7642±0.0346	0.3441±0.0170	0.5138±0.0533	0.7166±0.0081	0.6291±0.0559	0.8510±0.0113	**0.9713**±0.0000	0.7786±0.0006	0.9274±0.0007	0.9654±0.0016
SVHN	0.6169±0.0000	0.6382±0.0000	0.5899±0.0088	0.6170±0.0050	0.5392±0.0305	0.6075±0.0049	0.6137±0.0008	0.6213±0.0000	0.6291±0.0110	0.5993±0.0064	0.6244±0.0038	0.6445±0.0004	**0.6569**±0.0145
optdigits	0.9372±0.0000	0.9665±0.0000	0.8107±0.0328	0.9718±0.0080	0.3945±0.1865	0.6746±0.0494	0.9076±0.0212	0.9428±0.0167	0.8238±0.0268	0.8528±0.0000	0.9189±0.0079	0.8557±0.0010	**0.9955**±0.0004
CIFAR10	0.6753±0.0000	0.7030±0.0000	0.6404±0.0093	0.6360±0.0083	0.5612±0.0220	0.6753±0.0072	0.6791±0.0013	0.6782±0.0000	0.6853±0.0159	0.6453±0.0062	0.6554±0.0086	0.7098±0.0020	**0.7362**±0.0139
pageblocks	0.8965±0.0000	0.9130±0.0000	0.8264±0.0089	0.8839±0.0077	0.7839±0.0153	0.8805±0.0119	0.8693±0.0042	0.8932±0.0026	0.8989±0.0055	0.8658±0.0000	**0.9211**±0.0013	0.9088±0.0017	0.8051±0.0059
satimage-2	0.9971±0.0000	0.9938±0.0000	0.9912±0.0021	0.9948±0.0022	0.9294±0.0284	0.9899±0.0006	0.9962±0.0016	0.9967±0.0000	0.9934±0.0007	0.9840±0.0000	0.9668±0.0015	**0.9975**±0.0001	0.9060±0.0010
FashionMNIST	0.8987±0.0000	0.9160±0.0000	0.8415±0.0107	0.9056±0.0027	0.7545±0.0222	0.8803±0.0024	0.8852±0.0007	0.9014±0.0000	0.9021±0.0055	0.8666±0.0066	0.8981±0.0062	0.9121±0.0003	**0.9392**±0.0022
landsat	0.6825±0.0000	0.6658±0.0000	0.5880±0.0221	0.6513±0.0044	0.5944±0.0141	0.4052±0.0232	0.5137±0.0100	0.6820±0.0175	0.5279±0.0163	0.6953±0.0000	0.5905±0.0014	0.5209±0.0011	**0.7597**±0.0026
satellite	0.8224±0.0000	0.8030±0.0000	0.7746±0.0148	0.8515±0.0048	0.7619±0.0266	0.6876±0.0092	0.7770±0.0053	0.8211±0.0067	0.7861±0.0070	0.8498±0.0000	0.8002±0.0003	0.7377±0.0005	**0.8523**±0.0016
pendigits	0.9987±0.0000	0.9905±0.0000	0.9722±0.0048	0.9671±0.0083	0.4629±0.1135	0.8997±0.0203	0.9811±0.0023	0.9961±0.0017	0.9779±0.0068	0.9811±0.0000	0.9713±0.0014	0.9931±0.0003	**0.9999**±0.0000
annthyroid	0.9281±0.0000	0.8863±0.0000	0.9028±0.0152	0.8111±0.0107	0.5501±0.0362	0.8101±0.0516	0.8882±0.0134	0.9290±0.0025	**0.9752**±0.0015	0.8517±0.0000	0.9174±0.0003	0.9445±0.0013	0.8210±0.0218
mnist	0.8987±0.0000	0.9160±0.0000	0.8415±0.0107	0.9056±0.0027	0.7545±0.0222	0.8803±0.0024	0.8852±0.0007	0.9014±0.0000	0.9021±0.0055	0.8666±0.0066	0.8981±0.0062	0.9121±0.0003	**0.9927**±0.0003
MNIST-C	0.8411±0.0000	0.8721±0.0000	0.7675±0.0145	0.8562±0.0052	0.6470±0.0460	0.7934±0.0041	0.8012±0.0014	0.8474±0.0000	0.8610±0.0084	0.7728±0.0169	0.8662±0.0060	0.8565±0.0021	**0.9181**±0.0044
agnews	0.6705±0.0000	0.7458±0.0000	0.5843±0.0109	0.6255±0.0047	0.4983±0.0563	0.5987±0.0086	0.5782±0.0012	0.6798±0.0000	0.6816±0.0322	0.6096±0.0369	0.7443±0.0078	0.7220±0.0020	**0.8216**±0.0022
amazon	0.6058±0.0000	0.5788±0.0000	0.5640±0.0095	0.5421±0.0022	0.5120±0.0442	0.5607±0.0090	0.5509±0.0010	0.6082±0.0000	0.5671±0.0215	0.5474±0.0418	0.6152±0.0049	0.5945±0.0034	**0.6239**±0.0083
imdb	0.5008±0.0000	0.4957±0.0000	0.4953±0.0078	0.5234±0.0049	0.4997±0.0571	0.4846±0.0065	0.4791±0.0010	0.5043±0.0000	0.4805±0.0222	0.5084±0.0624	0.5106±0.0158	0.5011±0.0066	**0.5791**±0.0126
yelp	0.6807±0.0000	0.6720±0.0000	0.6107±0.0053	0.5579±0.0050	0.4990±0.0330	0.6107±0.0098	0.5930±0.0009	0.6866±0.0000	0.5994±0.0532	0.6029±0.0514	**0.6897**±0.0061	0.6715±0.0025	0.6638±0.0123
mammography	0.8758±0.0000	0.8552±0.0000	0.8802±0.0030	0.7187±0.0911	0.7150±0.0740	0.6994±0.0859	0.8101±0.0204	0.8762±0.0009	0.8642±0.0172	0.6911±0.0121	0.8924±0.0008	0.8224±0.0041	**0.9435**±0.0062
magic.gamma	0.8327±0.0000	0.8340±0.0000	0.7709±0.0129	0.7556±0.0042	0.6297±0.0107	0.6946±0.0239	0.8597±0.0108	0.8357±0.0076	0.8750±0.0090	0.8477±0.0014	0.8567±0.0005	0.8841±0.0008	**0.9355**±0.0005
campaign	0.7848±0.0000	0.7055±0.0000	0.7364±0.0148	0.8092±0.0079	0.6221±0.1288	0.4789±0.1232	0.7451±0.0042	0.7879±0.0025	0.7795±0.0111	0.6530±0.0070	0.7299±0.0059	0.7837±0.0021	**0.9217**±0.0012
shuttle	0.9991±0.0000	**0.9998**±0.0000	0.9965±0.0007	0.9992±0.0004	0.9979±0.0007	0.7044±0.1628	0.9991±0.0001	0.9993±0.0002	0.9975±0.0000	0.9982±0.0009	0.9993±0.0000	0.9992±0.0001	0.9996±0.0008
aloi	0.5104±0.0000	0.4876±0.0000	0.5074±0.0068	0.4750±0.0098	0.5089±0.0205	0.4801±0.0092	0.4991±0.0035	0.5119±0.0046	0.5044±0.0019	0.5405±0.0030	0.4987±0.0010	0.5304±0.0004	**0.6054**±0.0054
smtp	0.9243±0.0270	0.9342±0.0248	0.9036±0.0213	0.7436±0.0707	0.8524±0.0660	0.7878±0.1312	**0.9543**±0.0126	0.9298±0.0291	0.9527±0.0128	0.8304±0.0469	0.8411±0.0117	0.8967±0.0321	0.7698±0.0362
backdoor	0.9375±0.0053	0.9533±0.0025	0.7489±0.0267	0.9362±0.0069	0.9114±0.0262	0.5290±0.1448	0.8093±0.0056	0.9331±0.0170	0.9165±0.0173	0.7859±0.1711	0.8479±0.0409	0.9222±0.0056	**0.9903**±0.0011
celeba	0.7314±0.0072	0.4373±0.0075	0.7123±0.0227	0.7221±0.0082	0.5617±0.2246	0.4380±0.1049	0.7856±0.0199	0.7040±0.0037	0.8218±0.0238	0.7481±0.0061	0.7562±0.0052	0.6615±0.0065	**0.8847**±0.0126
skin	0.9949±0.0008	0.8634±0.0177	0.8942±0.0058	0.0658±0.0063	0.5995±0.0447	0.6495±0.0225	0.8874±0.0440	0.9886±0.0046	0.9177±0.0022	0.7060±0.0088	0.9057±0.0342	0.9701±0.0015	**0.9991**±0.0000
fraud	0.9543±0.0104	0.9435±0.0141	0.9473±0.0123	0.9278±0.0146	0.8313±0.0660	0.6975±0.2130	0.9365±0.0092	**0.9564**±0.0105	0.9352±0.0153	0.9391±0.0281	0.9100±0.0198	0.9546±0.0088	0.8961±0.0074
cover	0.9754±0.0015	0.9918±0.0010	0.8631±0.0207	0.8934±0.0402	0.4912±0.1474	0.1383±0.1334	0.9835±0.0066	0.9773±0.0055	0.9776±0.0128	0.9912±0.0029	0.6382±0.0434	**0.9939**±0.0012	0.9922±0.0017
census	0.7226±0.0029	0.5846±0.0106	0.6255±0.0238	0.7056±0.0035	0.5416±0.0433	0.3524±0.0419	0.7015±0.0023	0.7210±0.0040	0.6962±0.0091	0.6054±0.0496	0.6330±0.0362	0.6856±0.0152	**0.8475**±0.0059
http	1.0000±0.0000	0.9998±0.0003	0.9935±0.0029	0.9824±0.0345	0.6131±0.5149	0.9968±0.0013	1.0000±0.0000	0.9998±0.0003	0.9945±0.0010	0.9992±0.0005	0.9983±0.0008	1.0000±0.0000	1.0000±0.0000
donors	0.9949±0.0006	0.9697±0.0024	0.8944±0.0222	0.9990±0.0005	0.7295±0.1781	0.3357±0.1600	0.8250±0.0183	0.9926±0.0029	0.9815±0.0037	0.9991±0.0002	0.9519±0.0092	0.9882±0.0007	0.9991±0.0001
rAUC	0.907	0.884	0.851	0.895	0.769	0.779	0.875	0.915	0.902	0.899	0.910	0.925	**0.962**
WinRate	0.558	0.504	0.311	0.548	0.186	0.219	0.411	0.624	0.504	0.500	0.601	0.670	**0.798**
Avg. Rank	6.05	6.81	9.25	6.40	10.77	10.21	7.86	5.33	6.81	6.95	5.65	4.77	**3.37**