PURIST: POSITIVE–UNLABELED REFINEMENT VIA ITERATIVE SOFT-LABEL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Positive–unlabeled (PU) learning trains classifiers from trusted positives and mixed unlabeled examples. Existing methods compensate for missing negatives through different assumptions, but their effectiveness varies across data types and labeling conditions. We investigate which design principles transfer across heterogeneous PU problems. PURIST builds on two structural properties: under selected completely at random (SCAR) labeling, population-optimal P–U binary cross-entropy preserves optimal ranking at any fixed sampling ratio; label noise is confined to unlabeled examples. It combines balanced P:U optimization with Self-Adaptive EMA soft labels (SAEM), balancing training on trusted positives and noisy unlabeled examples while refining only unlabeled targets. Factorial ablations reveal a significant positive interaction between balancing and target refinement. Training dynamics further show that balancing improves target quality early, providing a more reliable starting point for subsequent temporal refinement. The method uses one model without a supplied class prior or data augmentation. Under selected at random (SAR) labeling, we analyze ranking preservation, AUC-regret bounds under anti-monotone selection, and non-identifiability under zero-correlation selection. Across 21 SCAR datasets from six source-data types, PURIST achieves an average AUC of 0.919 and the best average rank of 5.0 among 25 methods; it also ranks best on average under three proxy-defined SAR constructions on 14 datasets.

1 INTRODUCTION

Positive–unlabeled (PU) learning trains a binary classifier from labeled positives $\mathcal{P}$ and unlabeled examples $\mathcal{U}$ without confirmed negatives. This supervision arises in applications such as medical diagnosis (Chen et al., 2020b), fraud detection (Ortega Vázquez et al., 2022), and information retrieval (Jacovi et al., 2021). Classical PU methods are commonly organized into two-step approaches that mine reliable negatives, biased-learning approaches that treat unlabeled examples cost-sensitively, and class-prior-incorporation approaches that construct PU risks (Bekker & Davis, 2020). Following the PU-Bench taxonomy (Chen et al., 2026), modern deep methods span risk-minimization estimators such as nnPU (Kiryo et al., 2017) and Dist-PU (Zhao et al., 2022), disambiguation-guided empirical risk minimization (ERM) such as Self-PU (Chen et al., 2020b) and HolisticPU (Wang et al., 2023), and generative distribution matching such as VAE-PU (Na et al., 2020) and PAN (Hu et al., 2021).

These families compensate for missing negatives through assumptions with distinct failure modes. Prior-based risk estimators require the positive-class prior π, which is often difficult to determine in practice (du Plessis et al., 2015); two-step methods require reliable negatives (Liu et al., 2003); propensity models must separate the instance-dependent probability that a positive is labeled from the class posterior (Bekker et al., 2020; Gong et al., 2022; Gerych et al., 2022); and contrastive or generative methods add representation, augmentation, latent-variable, or adversarial structure (Long et al., 2024; Papič et al., 2023). A method effective under one data type or labeling mechanism may therefore fail as these conditions change.

PU-Bench finds that no method consistently wins across its image, text, and tabular benchmarks, with performance varying by data type, label frequency, and labeling mechanism (Chen et al., 2026). Complementary component studies disentangle pseudo-label targets, thresholding, mixup, and temporal averaging through controlled ablations, but remain confined to image benchmarks (Dai et al., 2025).

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.

Evidence therefore separates cross-data-type comparisons from component analysis within a single data type. The broader question is which PU design principles are necessary and remain effective across heterogeneous source data and labeling conditions.

Two structural properties suggest a simpler design. Under selected completely at random (SCAR) labeling, the labeled-to-unlabeled density ratio is proportional to the positive-class posterior $P(Y{=}1 \mid X{=}x)$ (Elkan & Noto, 2008); the posterior for distinguishing labeled positives from unlabeled examples is therefore strictly increasing in the class posterior and preserves the ranking relevant to area under the receiver operating characteristic curve (AUC) (Menon et al., 2015). Changing a fixed P:U sampling ratio alters only the monotone probability link of the population-optimal binary cross-entropy (BCE) scorer. Equal sampling therefore retains this ordering without knowing the class prior $\pi = P(Y{=}1)$, while controlling the relative influence of the trusted and noisy sides in finite-sample optimization (Sections 3.2 and 4.1). Second, PU noise is one-sided rather than symmetric: hidden positives in $\mathcal{U}$ are treated as negatives, while labels in $\mathcal{P}$ remain trustworthy (Elkan & Noto, 2008; Bekker & Davis, 2020). These observations separate the problem into two concrete needs: balance the trusted and noisy sides during optimization, and refine only the noisy $\mathcal{U}$ targets while preserving $\mathcal{P}$ as anchors.

PURIST implements this design directly. It combines balanced P:U sampling and standard BCE with SAEM (Self-Adaptive EMA soft labels), which progressively refines per-sample targets on $\mathcal{U}$ while leaving $\mathcal{P}$ fixed. The method uses a single model, requires no supplied class prior, explicit propensity model, or data augmentation, and shares optimizer and SAEM settings across datasets. In practical applications, the probability that a positive is labeled often varies with its features, making it important to understand method behavior under selected at random (SAR) labeling (Bekker et al., 2020). Yet analyses of ranking properties under SAR remain relatively limited in PU learning. Beyond SCAR, our analysis shows how SAR labeling changes the population ranking: monotone propensity preserves AUC optimality, anti-monotone propensity admits quantifiable degradation, and zero-correlation propensity leaves propensity and class posterior non-identifiable. To our knowledge, the SCAR experiments constitute the first principle-oriented controlled PU study spanning 21 datasets drawn from 6 source-data types; complete 26-method results are reported in the appendix. PURIST achieves the best average rank in the complete comparisons. We compare SAEM with alternative U-target strategies and find a significant positive interaction between P:U balancing and SAEM in ablation experiments.

2 Related Work

PU risk estimators requiring π. uPU (du Plessis et al., 2014; 2015) decomposes PU risk using a known prior π; nnPU (Kiryo et al., 2017) adds non-negative correction; batchwise truncation can bias gradients relative to the full-sample corrected risk (Section F). nnPUSB (Kato et al., 2019) extends this framework to selection bias while still requiring π and assumes invariance of order, a condition satisfied under SAR-A (Section 4.1). Self-PU (Chen et al., 2020b), Dist-PU (Zhao et al., 2022), PULDA (Jiang et al., 2023), PUET (Wilton et al., 2022), and VAE-PU (Na et al., 2020) require π in their training procedures, inheriting sensitivity to prior misspecification. RobustPU (Zhu et al., 2023) uses π during nnPU pretraining; its subsequent self-paced stage does not explicitly require π.

PU methods not requiring a supplied π. VPU (Chen et al., 2020a) uses a Donsker–Varadhan variational bound, avoiding π but requiring an irreducibility assumption. GradPU (Dai et al., 2023) combines an input-gradient penalty with positive upweighting to mitigate overfitting. LBE (Gong et al., 2022) jointly estimates the labeling propensity via dual-network expectation–maximization (EM) and therefore must separate propensity and class posterior from a likelihood that depends on their product. HolisticPU (Wang et al., 2023) also uses balanced resampling as a strong starting point, but handles training degradation through a two-stage pipeline (trend detection followed by hard partitioning and retraining); PURIST instead updates soft targets online without storing a full prediction trajectory or applying a separate trend-detection test. P3Mix (Li et al., 2022) applies mixup-based pseudo-labeling without π in its loss. BBE-PU (Garg et al., 2021) estimates π from data, with consistency guarantees under a high-score region of nonzero probability containing only positives. LaGAM (Long et al., 2024) combines contrastive learning with meta-label correction, and PAN (Hu et al., 2021) uses adversarial training. NcPU (Zhao et al., 2026) also fixes positive labels and iteratively updates unlabeled labels, but additionally uses a non-contrastive representation-alignment component.

Weakly supervised AUC optimization. WSAUC (Xie et al., 2024) casts AUC optimization under several forms of weak supervision, including PU learning, as learning from contaminated observed sets. It introduces reversed partial AUC (rpAUC) to reduce the influence of contaminated examples. Unlike methods that directly optimize pairwise AUC objectives, PURIST uses standard BCE with balanced sampling and soft-target refinement.

From semi-supervised and noisy-label learning to PU target refinement. Prediction-derived targets are widely used in semi-supervised learning: PseudoLabel converts predictions into hard targets (Lee, 2013), Temporal Ensembling averages per-sample predictions over time (Laine & Aila, 2017), and Mean Teacher instead averages model parameters to generate consistency targets (Tarvainen & Valpola, 2017). Related noisy-label methods use temporal targets to resist memorization: ELR regularizes predictions toward per-sample historical targets while retaining the observed-label cross-entropy, whereas SAT progressively replaces initialized targets using prediction EMA and confidence weighting (Liu et al., 2020; Huang et al., 2020).

PU learning combines missing labels with asymmetric one-sided noise: provisional negative targets in $\mathcal{U}$ may be wrong, while labels in $\mathcal{P}$ are trusted (Elkan & Noto, 2008; Bekker & Davis, 2020). Existing adaptations differ in which targets they update and how. SatPU updates targets across the training set with sharpening, reweighting, and optional temporal filtering (Song et al., 2025); LaGAM fixes $\mathcal{P}$ but derives $\mathcal{U}$ updates from projected meta-gradients coupled with contrastive learning (Long et al., 2024).

3 METHOD

3.1 PROBLEM SETTING

Let X be a random variable taking values in the input space $\mathcal{X}$, and let $Y \in \{0, 1\}$ denote its true class. Denote the positive- and negative-class distributions by p_+ and r, respectively. With the unknown positive-class prior $\pi = P(Y{=}1) \in (0, 1)$, the unlabeled distribution is $q = \pi p_+ + (1{-}\pi)r$.

In the case-control setting, we independently draw n_P labeled positives and n_U unlabeled examples to form $\mathcal{P}$ and $\mathcal{U}$, respectively, with unlabeled examples drawn from q (Elkan & Noto, 2008; Bekker & Davis, 2020). Let $S = 1$ indicate that an example is labeled, and define the positive-labeling propensity as $e(x) = P(S{=}1 \mid Y{=}1, X{=}x)$. Under SCAR, $e(x) = c > 0$, so labeled positives follow p_+. Under SAR, $e(x)$ may vary with x (Bekker & Davis, 2020; Bekker et al., 2020), and labeled positives instead follow

$$p_{\mathcal{P}}(x) = \frac{e(x)p_+(x)}{\mathbb{E}_{X\sim p_+}[e(X)]}.$$

We analyze the resulting ranking properties in Section 4.

Let $f_\theta : \mathcal{X} \to \mathbb{R}$ denote the scorer, with sigmoid output $\sigma(f_\theta(x))$. We evaluate its ranking performance by AUC. For independent $X_+ \sim p_+$ and $X_- \sim r$, define

$$\mathrm{AUC}(f; p_+, r) = \Pr\big(f(X_+) > f(X_-)\big) + \tfrac{1}{2}\Pr\big(f(X_+) = f(X_-)\big).$$

Ranking by the class posterior $\eta(x) = P(Y{=}1 \mid X{=}x)$ achieves the Bayes-optimal AUC, denoted by $\mathrm{AUC}^* = \mathrm{AUC}(\eta; p_+, r)$ (Menon et al., 2015). AUC values in prose and diagnostic plots use the $[0, 1]$ scale; result tables report $100\times\mathrm{AUC}$.

3.2 BALANCED STOCHASTIC BCE

Motivation. With $n_U \gg n_P$, natural-proportion training dilutes the trusted $\mathcal{P}$ signal. Under stochastic gradient descent (SGD), class imbalance can expose the minority class to higher directional gradient noise (Francazi et al., 2023). In PU learning this effect is especially damaging: $\mathcal{P}$ labels are certain while $\mathcal{U}$ labels are unreliable—they contain hidden positives treated as negatives. Imbalanced batches therefore let the *unreliable* side dominate gradient updates.

Design. Each training round combines the labeled positives with an equal-sized sample drawn from $\mathcal{U}$ without replacement (WOR), then optimizes standard BCE on randomly shuffled mini-batches. The low-label-rate fallback and epoch-level coverage rule are given with the full algorithm in Section F.

Population ranking. Under SCAR, class-probability estimation by P–U BCE preserves the Bayes-posterior order and hence AUC optimality (Menon et al., 2015). Any fixed P:U sampling prior induces the same ordering; equal sampling preserves it while addressing the finite-sample optimization imbalance above (see Section 4.1 for the formal statement and its SAR extensions).

Target refinement versus loss modification. PU risk estimators can introduce nonlinear batch-level operations: nnPU truncates its empirical negative-risk term with $\max(0, \cdot)$ (Kiryo et al., 2017), whereas VPU applies a logarithm to the mini-batch mean of unlabeled predictions (Chen et al., 2020a). Consequently, their stochastic losses depend nonlinearly on the sampled unlabeled batch. Noise-robust methods such as ELR (Liu et al., 2020) retain the observed labels in the primary cross-entropy term and control their influence through an additional regularizer toward early predictions. PURIST instead leaves the BCE form unchanged and directly refines its training targets on $\mathcal{U}$, while keeping the reliable $\mathcal{P}$ targets fixed. For a fixed pointwise scorer and frozen targets, uniform mini-batch sampling gives unbiased gradients of the BCE sampling objective.

3.3 SAEM: Self-Adaptive EMA Soft Labels

Motivation. Balanced BCE with hard labels $y_j = 0$ for all $x_j \in \mathcal{U}$ trains hidden positives against negative targets. PU learning has a one-sided label-noise structure: hidden positives in $\mathcal{U}$ are treated as negatives, while labels in $\mathcal{P}$ remain trustworthy (Elkan & Noto, 2008; Bekker & Davis, 2020). We therefore progressively refine the targets on $\mathcal{U}$ while keeping $\mathcal{P}$ fixed, without external assumptions about π or the labeling propensity.

Design. Building on per-example temporal prediction targets in semi-supervised and noisy-label learning (Laine & Aila, 2017; Huang et al., 2020), SAEM specializes this mechanism to PU supervision by updating only $\mathcal{U}$ and keeping $\mathcal{P}$ fixed. We maintain soft labels $\varphi_j \in [0, 1]$ for each $x_j \in \mathcal{U}$, initialized to 0. After each epoch, soft labels are updated via exponential moving average:

$$\varphi_j \leftarrow (1 - \gamma_t)\,\varphi_j + \gamma_t\,\sigma(f_\theta(x_j)), \quad \gamma_t = \max\big(\gamma_{\min},\ \tfrac{2}{t+2}\big), \qquad 0 < \gamma_{\min} \le 1. \tag{1}$$

Positive labels are never modified: $y_i = 1$ for all $x_i \in \mathcal{P}$ throughout training. The training loss uses φ as BCE targets for $\mathcal{U}$:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{x_i \in \mathcal{P}} \ell_{\mathrm{BCE}}(f_\theta(x_i), 1) + \frac{1}{|\mathcal{U}_{\mathrm{batch}}|} \sum_{x_j \in \mathcal{U}_{\mathrm{batch}}} \ell_{\mathrm{BCE}}(f_\theta(x_j), \varphi_j). \tag{2}$$

Temporal target refinement. SAEM is a direct output-space temporal smoothing rule. At the end of each epoch, the current model predictions refresh the per-sample targets on $\mathcal{U}$; the next epoch then optimizes standard BCE against those refreshed targets. The update uses one model and leaves the trusted targets on $\mathcal{P}$ unchanged.

Design points. The schedule starts with the decaying term $2/(t+2)$ and is lower-bounded by a positive $\gamma_{\min}$. The decaying term dampens the influence of individual predictions after the first epochs, while the positive floor prevents φ from freezing at early—potentially poor—predictions. Under exact alternating population updates, the resulting finite-time scorer preserves the balanced-BCE ranking, whereas indefinite updates compress score differences; Section 4.2 gives the formal result. When defined, validation proxy AUC selects the checkpoint; otherwise, training uses a fixed budget (Section F).

Training alternates the target refresh in Equation (1) with balanced BCE updates in Equation (2); the complete procedure and all implementation hyperparameters are reported in Section F. SAEM maintains one soft target per unlabeled example, requiring $O(|\mathcal{U}|)$ additional storage; further scaling details appear in Section J.

4 Theoretical Analysis

We state the population-level properties that motivate PURIST's design, assuming a function class with sufficient capacity.

4.1 Population Ranking under Fixed-Ratio P–U BCE

Class-probability estimation with BCE can optimize AUC under mutually contaminated labels because its population-optimal score preserves the relevant ordering (Menon et al., 2015). For P–U classification, let $\alpha \in (0,1)$ be any fixed sampling prior assigned to $\mathcal{P}$. Under SCAR, the population-optimal BCE probability is

$$w_\alpha(x) = \frac{\alpha p_+(x)}{\alpha p_+(x) + (1-\alpha)q(x)} = \frac{\alpha\eta(x)}{\alpha\eta(x) + (1-\alpha)\pi}.$$

It is strictly increasing in the Bayes posterior $\eta(x) = P(Y{=}1 \mid x)$ for every fixed α, and therefore induces a Bayes-optimal AUC ranking without knowledge of π. Equal P:U sampling sets $\alpha = 1/2$ and retains this ordering; its separate finite-sample role is to give the two sides equal objective weight. The corresponding SCAR density-ratio relation also follows from Elkan & Noto (2008).

Most PU theory assumes SCAR (Selected Completely At Random): every positive has the same probability of being labeled (Elkan & Noto, 2008; Bekker & Davis, 2020). SAR (Selected At Random) instead lets the labeling propensity $e(x) = P(S{=}1 \mid Y{=}1, X{=}x)$ vary with x (Bekker et al., 2020). We next analyze fixed-ratio P–U BCE under three controlled relations between propensity and posterior—monotone, anti-monotone, and zero-correlation—assuming throughout that $e(x) > 0$ for p_+-almost every x. Under SAR,

$$\frac{p_{\mathcal{P}}(x)}{q(x)} = \frac{e(x)\eta(x)}{\mathbb{E}_{X\sim q}[e(X)\eta(X)]},$$

and, for every fixed α, the BCE optimum $w^*_{e,\alpha}(x) = \alpha p_{\mathcal{P}}(x)/[\alpha p_{\mathcal{P}}(x) + (1-\alpha)q(x)]$ is a strictly increasing transform of this density ratio. Thus the sampling prior changes the output link but not the ordering; selection affects population AUC through the order induced by $e(x)\eta(x)$. Here $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ denotes total-variation distance, and Corr denotes Pearson correlation. Let $w^*_{e,\alpha} : \mathcal{X} \to [0,1]$ denote this population-optimal probability output, defined q-almost everywhere:

Proposition 4.1 (Fixed-ratio P–U BCE ranking under SAR)**.** *Fix any sampling prior $\alpha \in (0,1)$ for $\mathcal{P}$.*

1. ***SAR-A*** *(monotone: e nondecreasing in η; Gong et al., 2022).* $\mathrm{AUC}(w^*_{e,\alpha}) = \mathrm{AUC}^*$*; SCAR is a special case.*

2. ***SAR-B*** *(anti-monotone: e decreasing in η; Gong et al., 2022). The population-optimal probability ranks by $e(x)\eta(x)$, and*

$$\mathrm{AUC}^* - \mathrm{AUC}(w^*_{e,\alpha}) \le \min\left\{ \frac{2}{1-\pi}\,\mathrm{TV}(p_{\mathcal{P}}, p_+), 2\,\mathrm{AUC}^* - 1 \right\}.$$

   The two terms respectively control the total selection-induced shift and the intrinsic rank signal, while the exact pairwise decomposition in Section A counts only posterior pairs whose $e\eta$ order is reversed or tied. The bound holds for any SAR propensity and vanishes under SCAR. Anti-monotonicity alone does not imply degradation: the ranking remains optimal whenever $\eta(x_1) > \eta(x_2)$ implies $e(x_1)\eta(x_1) > e(x_2)\eta(x_2)$ for $q \otimes q$-almost every pair.

3. ***SAR-C*** *(orthogonal: both variables have positive variance and* $\mathrm{Corr}_{X\sim q}(e(X), \eta(X)) = 0$*). The observed distribution identifies only the normalized product $e\eta$; zero correlation does not separate the two factors.*

Under general SAR, separating propensity from class posterior requires additional distributional or parametric assumptions (Gerych et al., 2022; Furmańczyk et al., 2023).

4.2 Finite-Time Rank Preservation of SAEM

The preceding result characterizes the initial hard-target scorer. We next consider an idealized alternating population update over measurable probability scores that also captures SAEM's evolving targets. Let $w(x) = p_{\mathcal{P}}(x)/(p_{\mathcal{P}}(x) + q(x))$ be the population-optimal balanced-BCE probability when the unlabeled target is zero. Starting from $\phi_0(x) = 0$, let q_t exactly minimize population balanced BCE for the current unlabeled target ϕ_t, then update $\phi_{t+1} = (1-\gamma_{t+1})\phi_t + \gamma_{t+1}q_t$.

Table 1: Test AUC ($\times 100$) on 21 datasets (SCAR, c=0.1, mean±std over 10 seeds except where indicated). †: PN oracle. Best displayed non-oracle in **bold**. DT is the balanced-BCE-only ablation (PURIST without SAEM). Ranks are computed against all 25 evaluated non-oracle methods. Groups: Oracle | ours+controlled U-label | generic noisy-label | PU(π-dep.) | PU(π-free).

Dataset	Type	PN†	**PURIST**	PseudoLabel	ELR	DT	MeanTeacher	SAT	nnPUSB	WSAUC	PUET	Dist-PU	PULDA	Holistic	Direct-AUC	LaGAM	VPU
MNIST	Image	$99.8_{\pm 0.1}$	$99.5_{\pm 0.1}$	$99.6_{\pm 0.1}$	$99.3_{\pm 0.5}$	$99.1_{\pm 0.7}$	$98.5_{\pm 1.3}$	$99.3_{\pm 0.3}$	$95.2_{\pm 3.2}$	$99.1_{\pm 1.5}$	$99.2_{\pm 0.1}$	$\mathbf{99.8}_{\pm 0.0}$	$98.7_{\pm 2.2}$	$98.7_{\pm 0.9}$	$98.5_{\pm 1.3}$	$97.6_{\pm 1.0}$	$99.0_{\pm 0.8}$
FashionMNIST	Image	$99.9_{\pm 0.1}$	$99.6_{\pm 0.0}$	$99.6_{\pm 0.1}$	$99.5_{\pm 0.2}$	$99.2_{\pm 0.4}$	$98.4_{\pm 2.2}$	$99.5_{\pm 0.1}$	$98.6_{\pm 0.7}$	$99.4_{\pm 0.6}$	$\mathbf{99.7}_{\pm 0.0}$	$99.4_{\pm 0.3}$	$99.1_{\pm 0.6}$	$99.5_{\pm 0.3}$	$98.9_{\pm 0.6}$	$99.1_{\pm 0.4}$	$99.0_{\pm 0.9}$
CIFAR10	Image	$98.5_{\pm 0.6}$	$95.6_{\pm 0.4}$	$96.0_{\pm 1.2}$	$94.9_{\pm 1.1}$	$94.0_{\pm 3.0}$	$95.2_{\pm 0.8}$	$94.9_{\pm 1.1}$	$95.7_{\pm 0.5}$	$95.6_{\pm 1.4}$	$88.0_{\pm 0.2}$	$96.7_{\pm 0.6}$	$96.4_{\pm 0.9}$	$97.1_{\pm 0.7}$	$94.2_{\pm 1.7}$	$\mathbf{97.6}_{\pm 1.3}$	$94.3_{\pm 2.2}$
CIFAR100	Image	$95.1_{\pm 0.9}$[5]	$79.0_{\pm 6.2}$[5]	$78.9_{\pm 1.4}$[5]	$85.2_{\pm 2.6}$[5]	$77.5_{\pm 3.7}$[5]	$61.8_{\pm 16.4}$[5]	$81.4_{\pm 4.4}$[5]	$82.1_{\pm 2.8}$[5]	$86.6_{\pm 1.5}$[5]	$70.3_{\pm 1.4}$[5]	$\mathbf{87.0}_{\pm 1.6}$[5]	$86.0_{\pm 2.7}$[5]	$78.5_{\pm 18.1}$[5]	$81.7_{\pm 2.6}$[5]	$86.0_{\pm 1.5}$[5]	$82.0_{\pm 4.7}$[5]
SVHN	Image	$98.7_{\pm 0.8}$	$89.2_{\pm 13.2}$	$59.6_{\pm 8.3}$	$93.8_{\pm 2.0}$	$91.9_{\pm 3.3}$	$90.1_{\pm 4.1}$	$94.4_{\pm 1.0}$	$93.3_{\pm 1.8}$	$95.4_{\pm 0.9}$	$75.1_{\pm 0.9}$	$50.0_{\pm 0.0}$	$53.3_{\pm 3.4}$	$\mathbf{97.3}_{\pm 0.4}$	$93.4_{\pm 0.8}$	$89.8_{\pm 2.6}$	$88.0_{\pm 1.2}$
Adult	Tabular	$91.2_{\pm 0.2}$	$87.7_{\pm 0.6}$	$88.4_{\pm 0.5}$	$87.5_{\pm 1.2}$	$86.8_{\pm 2.0}$	$87.2_{\pm 2.5}$	$90.1_{\pm 0.5}$	$89.3_{\pm 1.3}$	$87.0_{\pm 1.0}$	$85.9_{\pm 0.5}$	$89.2_{\pm 0.6}$	$88.3_{\pm 1.3}$	$89.1_{\pm 1.5}$	$86.0_{\pm 1.6}$	$88.6_{\pm 0.7}$	$\mathbf{90.2}_{\pm 0.6}$
Covertype	Tabular	$97.5_{\pm 0.5}$	$91.3_{\pm 0.5}$	$90.0_{\pm 0.4}$	$91.1_{\pm 0.7}$	$92.8_{\pm 1.2}$	$91.2_{\pm 0.3}$	$85.5_{\pm 0.3}$	$92.2_{\pm 0.3}$	$90.7_{\pm 1.9}$	$\mathbf{95.1}_{\pm 0.1}$	$86.3_{\pm 0.3}$	$81.3_{\pm 1.2}$	$88.6_{\pm 0.5}$	$93.8_{\pm 0.4}$	$89.2_{\pm 0.4}$	$83.3_{\pm 0.2}$
SUSY	Tabular	$87.5_{\pm 0.2}$	$\mathbf{87.0}_{\pm 0.1}$	$86.8_{\pm 0.2}$	$86.0_{\pm 0.3}$	$86.9_{\pm 0.1}$	$87.0_{\pm 0.2}$	$86.3_{\pm 0.2}$	$86.8_{\pm 0.1}$	$84.7_{\pm 0.7}$	$85.3_{\pm 0.1}$	$85.7_{\pm 0.3}$	$83.0_{\pm 0.7}$	$86.8_{\pm 0.3}$	$86.6_{\pm 0.3}$	$86.4_{\pm 0.2}$	$86.0_{\pm 0.3}$
Pendigits	Tabular	$99.8_{\pm 0.2}$	$\mathbf{99.7}_{\pm 0.2}$	$99.2_{\pm 0.3}$	$99.4_{\pm 0.5}$	$99.6_{\pm 0.1}$	$99.4_{\pm 0.3}$	$96.7_{\pm 4.3}$	$96.0_{\pm 3.8}$	$97.0_{\pm 3.4}$	$99.4_{\pm 0.2}$	$99.2_{\pm 0.2}$	$98.7_{\pm 1.5}$	$99.4_{\pm 0.3}$	$99.4_{\pm 0.3}$	$97.8_{\pm 1.9}$	$95.0_{\pm 4.1}$
Letter	Tabular	$98.4_{\pm 2.4}$	$96.0_{\pm 0.7}$	$95.4_{\pm 0.7}$	$94.0_{\pm 2.5}$	$95.6_{\pm 0.5}$	$94.6_{\pm 0.9}$	$90.8_{\pm 5.6}$	$90.5_{\pm 8.0}$	$95.7_{\pm 1.8}$	$94.8_{\pm 0.8}$	$89.7_{\pm 2.4}$	$\mathbf{96.3}_{\pm 0.9}$	$95.7_{\pm 0.9}$	$95.1_{\pm 1.0}$	$89.6_{\pm 3.4}$	$89.5_{\pm 4.3}$
Spambase	Tabular	$96.4_{\pm 1.1}$	$93.4_{\pm 1.2}$	$93.4_{\pm 0.8}$	$93.0_{\pm 2.1}$	$89.3_{\pm 3.4}$	$88.5_{\pm 6.0}$	$82.4_{\pm 16.7}$	$82.9_{\pm 9.3}$	$92.0_{\pm 0.9}$	$\mathbf{95.9}_{\pm 0.6}$	$93.5_{\pm 1.4}$	$84.7_{\pm 14.2}$	$90.3_{\pm 3.0}$	$90.0_{\pm 3.0}$	$90.1_{\pm 4.5}$	$87.4_{\pm 5.4}$
Mushrooms	Tabular	$100.0_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.6_{\pm 0.5}$	$99.5_{\pm 0.6}$	$99.9_{\pm 0.2}$	$99.9_{\pm 0.1}$	$99.7_{\pm 0.3}$	$\mathbf{100.0}_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.4_{\pm 1.7}$	$99.0_{\pm 1.2}$	$99.9_{\pm 0.1}$	$99.4_{\pm 1.0}$	$99.2_{\pm 1.6}$
Connect4	Tabular	$95.4_{\pm 1.1}$	$87.8_{\pm 0.7}$	$86.8_{\pm 0.8}$	$85.8_{\pm 1.6}$	$85.4_{\pm 1.1}$	$86.2_{\pm 1.0}$	$76.9_{\pm 7.9}$	$86.1_{\pm 1.3}$	$83.9_{\pm 5.6}$	$82.8_{\pm 0.7}$	$70.5_{\pm 5.2}$	$85.0_{\pm 1.3}$	$83.2_{\pm 5.5}$	$85.5_{\pm 1.3}$	$82.5_{\pm 2.0}$	$\mathbf{87.9}_{\pm 1.0}$
IMDB	Text	$88.7_{\pm 0.8}$	$83.2_{\pm 1.2}$	$83.5_{\pm 0.9}$	$79.8_{\pm 3.3}$	$77.8_{\pm 4.4}$	$79.4_{\pm 6.5}$	$62.7_{\pm 7.0}$	$\mathbf{84.2}_{\pm 1.4}$	$80.2_{\pm 2.9}$	$75.5_{\pm 0.5}$	$56.7_{\pm 13.4}$	$81.8_{\pm 3.6}$	$83.6_{\pm 3.1}$	$78.2_{\pm 2.8}$	$82.7_{\pm 1.4}$	$80.6_{\pm 7.3}$
20News	Text	$96.6_{\pm 0.9}$	$91.3_{\pm 1.1}$	$91.4_{\pm 1.0}$	$90.8_{\pm 1.8}$	$89.0_{\pm 1.8}$	$79.0_{\pm 11.5}$	$65.9_{\pm 11.1}$	$92.2_{\pm 0.6}$	$88.8_{\pm 3.7}$	$87.9_{\pm 0.6}$	$\mathbf{92.6}_{\pm 1.1}$	$86.8_{\pm 10.5}$	$87.9_{\pm 6.9}$	$88.9_{\pm 2.0}$	$89.9_{\pm 6.6}$	$89.6_{\pm 2.6}$
DBpedia	Text	$99.7_{\pm 0.1}$	$98.7_{\pm 0.2}$	$98.5_{\pm 0.3}$	$97.3_{\pm 1.5}$	$98.0_{\pm 0.3}$	$84.5_{\pm 15.4}$	$95.1_{\pm 6.7}$	$98.9_{\pm 0.2}$	$98.8_{\pm 0.2}$	$93.1_{\pm 0.3}$	$99.3_{\pm 0.2}$	$\mathbf{99.4}_{\pm 0.3}$	$98.3_{\pm 0.7}$	$97.9_{\pm 0.4}$	$99.2_{\pm 0.1}$	$98.4_{\pm 1.1}$
Elliptic	Graph	$98.4_{\pm 0.6}$	$96.1_{\pm 0.3}$	$96.1_{\pm 0.4}$	$95.9_{\pm 0.8}$	$95.4_{\pm 0.7}$	$92.6_{\pm 1.7}$	$95.3_{\pm 0.5}$	$96.0_{\pm 0.7}$	$96.2_{\pm 0.4}$	$\mathbf{96.6}_{\pm 0.5}$	$94.3_{\pm 0.6}$	$95.8_{\pm 0.6}$	$88.6_{\pm 13.3}$	$95.2_{\pm 0.7}$	$94.3_{\pm 0.8}$	$93.4_{\pm 1.4}$
DiFraud	Text	$89.5_{\pm 0.5}$	$\mathbf{83.4}_{\pm 0.7}$	$82.5_{\pm 0.7}$	$79.9_{\pm 1.2}$	$80.1_{\pm 2.1}$	$78.6_{\pm 2.8}$	$63.4_{\pm 4.5}$	$82.6_{\pm 0.4}$	$82.2_{\pm 0.6}$	$77.2_{\pm 0.3}$	$66.5_{\pm 16.5}$	$81.3_{\pm 1.2}$	$82.4_{\pm 0.8}$	$80.1_{\pm 1.8}$	$79.7_{\pm 0.7}$	$81.2_{\pm 1.6}$
SpeechCommands	Audio	$97.9_{\pm 0.4}$	$92.6_{\pm 1.2}$	$63.6_{\pm 17.1}$	$95.1_{\pm 0.8}$	$88.0_{\pm 4.4}$	$89.5_{\pm 3.2}$	$93.4_{\pm 1.4}$	$93.3_{\pm 1.2}$	$\mathbf{96.3}_{\pm 0.3}$	$75.1_{\pm 1.0}$	$51.7_{\pm 1.2}$	$89.8_{\pm 12.7}$	$53.9_{\pm 11.8}$	$94.5_{\pm 1.0}$	$86.4_{\pm 5.6}$	$92.4_{\pm 0.6}$
YelpChi	Graph	$87.3_{\pm 3.1}$	$80.1_{\pm 0.5}$	$79.6_{\pm 1.5}$	$79.1_{\pm 1.9}$	$76.9_{\pm 3.8}$	$76.5_{\pm 3.6}$	$79.7_{\pm 1.7}$	$78.0_{\pm 5.1}$	$78.9_{\pm 1.5}$	$\mathbf{80.6}_{\pm 0.7}$	$79.2_{\pm 0.6}$	$78.1_{\pm 1.3}$	$74.7_{\pm 2.5}$	$79.3_{\pm 2.8}$	$76.7_{\pm 1.4}$	$73.1_{\pm 4.0}$
CICIDS2017	Network	$99.7_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.5_{\pm 0.1}$	$99.4_{\pm 0.1}$	$99.5_{\pm 0.0}$	$99.4_{\pm 0.0}$	$98.3_{\pm 0.1}$	$99.4_{\pm 0.0}$	$99.3_{\pm 0.2}$	$\mathbf{99.9}_{\pm 0.0}$	$98.8_{\pm 0.1}$	$98.7_{\pm 0.1}$	$99.1_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.3_{\pm 0.1}$	$98.0_{\pm 0.0}$
Avg		96.0	91.9	89.0	91.7	90.6	88.4	87.2	91.1	91.8	88.4	85.0	88.7	89.1	91.3	90.6	89.9
Rank		—	5.0	6.9	9.0	10.6	13.4	12.6	8.7	9.0	10.8	11.0	12.0	10.4	10.5	11.6	12.8

Proposition 4.2 (Finite-time SAEM rank preservation)**.** *If $0 < \gamma_t \leq 1$, then for every finite t, q_t is strictly increasing in w; ϕ_t is also strictly increasing for $t \geq 1$. Thus idealized SAEM preserves the ranking of its balanced-BCE starting scorer at every finite epoch.*

Indeed, pointwise BCE minimization and the target update give $q_t = w + (1-w)\phi_t$ and $\phi_t = 1 - \prod_{s=1}^{t}(1-\gamma_s w)$, respectively. These expressions make q_t strictly increasing in w for every finite t, and ϕ_t strictly increasing for $t \geq 1$. Consequently, the finite-time scorer retains the Bayes order under SCAR, SAR-A, and SAR-B satisfying the order condition in Proposition 4.1, while under general SAR it inherits the $e\eta$ order characterized there.

If $\sum_t \gamma_t = \infty$, then $\phi_t(x), q_t(x) \to 1$ wherever $w(x) > 0$, compressing score differences in the infinite-time limit. This limiting behavior motivates checkpoint selection in finite-sample training; our implementation uses validation proxy AUC. The proposition concerns exact population updates; finite-sample target gradient-coefficient distortion and SGD trajectories are examined in Section I. Proof in Section A.1.

5 Experiments

5.1 Setup

We build on the PU-Bench framework (Chen et al., 2026). The SCAR panel contains 21 binary tasks with reproducible preprocessing and a common case-control protocol, spanning 6 source-data types (image, tabular, text, graph, audio, network traffic). We use labeling rate c=0.1, where c is the fraction of positives in the fully labeled source split sampled into $\mathcal{P}$ (Chen et al., 2026), and 10 random seeds per configuration, except CIFAR100 with 5 seeds. Dataset-level cells report mean±standard deviation over seeds; summary tables state their cross-dataset aggregation in their captions. PURIST shares optimizer and SAEM settings across datasets, with backbone-specific batch sizes given in Section F. The controlled U-label baselines share PURIST's backbone, optimizer, balanced sampling, and early-stopping configuration and differ only in their U-label mechanism. Other PU baselines retain method-specific training structure and use source dataset recipes where available or fixed method-specific benchmark presets. Architectures, selection policy, and baseline configurations and adaptations are reported in Sections C and G. Checkpoint selection uses validation proxy AUC when defined, measuring separation of observed $\mathcal{P}$ from $\mathcal{U}$ on the validation split (Chen et al., 2026). Under SCAR, the population-optimal proxy ranking agrees with the class posterior; under SAR, it follows $e(x)\eta(x)$. Alignment with the class-posterior ranking cannot generally be guaranteed from PU observations alone because propensity and posterior are not identifiable. Fixed-budget exceptions and further details appear in Sections F and G.

The SCAR comparison contains a fully supervised positive–negative (PN) oracle and 25 non-oracle methods: five controlled U-label methods, SAT, nine PU methods whose evaluated implementations receive the true class prior π, and ten that do not. The complete inventory and adaptations appear in Section G. The body table displays all controlled variants, SAT, and nine external PU baselines spanning both prior-input groups and multiple estimator/disambiguation mechanisms. Its rank row uses the complete set of 25 non-oracle methods; their per-dataset results are reported in Section D.

5.2 Main Results (SCAR)

Table 1 shows test AUC on all 21 datasets. The best displayed method varies across datasets and method families. PURIST leads both aggregate summaries: its complete-comparison average rank is 5.0 versus 6.9 for the next-ranked PseudoLabel, and its average AUC is 0.919 versus 0.918 for the next-highest displayed non-oracle method, WSAUC. Full results for all 26 evaluated methods appear in Section D.

5.3 Controlled Comparison of U-Label Strategies

SAEM shares temporal target estimation with methods from both noisy-label learning and semi-supervised learning. To isolate the U-label strategy from shared training infrastructure, we run a controlled comparison where all four methods share *identical* backbone, optimizer, sampling, and early stopping; only the U-label mechanism differs. Hyperparameter details appear in Section G.

Table 2: Controlled comparison of U-label strategies. Rank is computed among the four methods.

U-label strategy	Avg Rank	Avg AUC	#Best	#Worst
SAEM (PURIST)	**1.52**	**91.9**	**11**	**0**
PseudoLabel	2.14	89.0	7	4
ELR	2.95	91.7	3	5
MeanTeacher	3.38	88.4	0	12

The mechanisms expose $\mathcal{U}$ to different targets: PseudoLabel converts predictions into hard targets, ELR retains the observed targets in its primary BCE and adds a temporal-target regularizer, and Mean Teacher obtains consistency targets from a parameter-EMA teacher. SAEM directly replaces every $\mathcal{U}$ target with its temporal soft label.

SAEM achieves the best average rank (1.52 among four), ranks best on 11 of the 21 datasets, and is never last among the four strategies. Mean Teacher has average rank 3.38; its parameter-space EMA changes prediction targets indirectly through the teacher model, whereas SAEM updates the output-space training targets directly.

To examine the balance–refinement interaction, we hold SAEM fixed and vary only P:U sampling. Using the latent U classes for diagnosis, the dataset-averaged trajectories in Figure 1 show separation between hidden-positive and true-negative targets from the first refresh under balanced sampling, whereas both remain closer to zero under natural proportions.

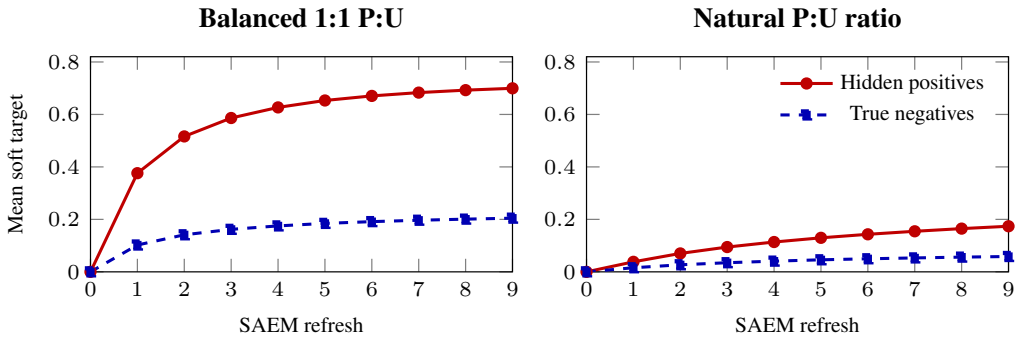

Figure 1: SAEM U-target trajectories under balanced and natural-proportion P:U sampling. Curves show mean soft targets for latent hidden positives and true negatives. Each curve equally averages 21 dataset means, each computed from ten paired seeds. All targets start at zero.

5.4 SAR Robustness

To examine departures from SCAR, we follow the SAR-A/B regimes of Gong et al. (2022) using a PN-trained posterior proxy $s(x)$. These proxy-defined constructions use s in place of the latent posterior η. SAR-A uses increasing proxy weights, whereas SAR-B uses weights $(2.5 - s(x))^{10}$. At the weight level, the product $s(2.5 - s)^{10}$ decreases above $s = 2.5/11$ (Proposition 4.1), favoring positives with lower proxy scores. SAR-C residualizes a feature-centroid proximity score against s over the source population and maps the residual to positive inclusion propensities, making the

Table 3: Average AUC ($\times 100$) and rank under SCAR (21 datasets) and proxy-defined SAR (14). Ranks use all 25 non-oracle methods; AUC reports mean$\pm$std across dataset-level seed means. Best average rank in **bold**; full results in Section D.

	SCAR		SAR-A		SAR-B		SAR-C	
Method	AUC	Rank	AUC	Rank	AUC	Rank	AUC	Rank
PURIST	91.9±6.6	**5.0**	92.0±6.7	**4.4**	79.2±21.7	**6.6**	93.0±6.5	**4.1**
PseudoLabel	89.0±11.1	6.9	91.7±6.8	7.6	74.7±23.4	8.9	91.5±8.0	6.3
nnPUSB	91.1±6.4	8.7	90.9±6.7	10.4	74.7±23.0	12.4	91.2±7.2	10.4
WSAUC	91.8±6.7	9.0	91.6±7.3	9.6	77.0±22.9	10.5	92.2±6.9	9.9
ELR	91.7±6.7	9.0	90.9±6.8	11.5	78.0±21.5	9.4	92.1±6.9	9.5
Holistic	89.1±10.8	10.4	86.9±13.8	11.9	73.2±24.9	12.6	88.7±12.8	10.4
Direct-AUC	91.3±7.0	10.5	91.3±7.2	10.3	79.6±20.3	6.7	91.9±7.2	11.0
DT	90.6±7.6	10.6	91.3±7.0	9.9	78.6±20.5	8.1	91.5±7.0	12.1
Dist-PU	85.0±16.0	11.0	88.8±12.6	8.9	71.9±25.1	13.4	87.4±14.5	10.4
LaGAM	90.6±6.8	11.6	90.6±7.8	13.1	79.9±22.0	7.9	91.2±7.6	12.4
PULDA	88.7±10.7	12.0	90.5±7.7	9.9	78.5±23.4	8.4	90.9±8.5	10.3
SAT	87.2±11.6	12.6	89.5±7.5	12.4	75.6±23.9	11.7	89.7±10.0	12.8
VPU	89.9±7.0	12.8	89.4±7.9	16.1	76.0±21.9	14.2	91.4±7.5	12.6
MeanTeacher	88.4±9.3	13.4	89.3±6.9	13.8	76.0±19.4	13.5	90.5±6.5	14.3
LBE	84.2±12.9	15.0	87.6±10.0	11.8	71.8±24.0	13.4	86.1±13.2	13.1

prescribed propensity orthogonal to the posterior proxy. Exact sampling rules are given in Section B. Proxy quality, source-label alignment, and an exact known-η illustration are reported in Section B.1.

PURIST achieves the best average rank under SCAR and all three proxy-defined SAR constructions (Table 3), including SAR-B, which preferentially labels positives with low proxy scores. LBE (Gong et al., 2022) provides a relevant comparison through its dual-network EM formulation, which jointly estimates the class posterior and labeling propensity. Under SAR-C, PURIST achieves an average AUC of 0.930, compared with 0.861 for LBE, demonstrating strong ranking performance in this construction without explicit propensity estimation. Full results on 14 datasets with ten seeds per configuration appear in Section D.

5.5 Ablation

We conduct a 2×2 component ablation of P:U balancing and SAEM target refinement. Unbalanced variants use all available P and U samples in their natural proportions, while balanced variants follow PURIST's P:U sampling rule. Under each sampling condition, we compare hard targets for U with SAEM-refined targets, using the same refinement schedule for both SAEM variants. All four variants select checkpoints using the same validation proxy AUC criterion.

Table 4: Two-by-two component ablation (SCAR, c=0.1). AUC is averaged over seeds within each dataset, then equally across 21 datasets. Reported p-values come from two-sided Wilcoxon signed-rank tests on paired dataset means, Holm-adjusted across six contrasts. All six effects and their 95% confidence intervals appear in Table 21.

P:U balancing	Hard U	SAEM	ΔAUC (pp)	p
No	90.1	90.1	0.00	0.514
Yes	90.6	**91.9**	+1.33	0.010
Balance × SAEM interaction			+1.34	0.014
Joint method vs. unbalanced BCE			+1.87	0.014

SAEM has negligible effect on average AUC under natural P:U sampling and gains 1.33 percentage points under balanced sampling (0.906 to 0.919). The latter gain, the balance–refinement interaction, and the joint method's gain over unbalanced hard-target BCE are statistically significant (Table 4). Neither the SAEM effect under natural sampling nor the balancing effect with hard targets is significant.

At the first SAEM refresh, squared U-target error against the latent labels falls by 50.8% under balanced sampling and 7.3% under natural sampling, relative to the initial all-zero targets. These reductions average ten paired seeds within each of 21 equally weighted datasets. This error equals the squared discrepancy in BCE logit-gradient coefficients. Balanced sampling yields larger reductions in 203/210 seed pairs and all 21 dataset means (dataset-level Wilcoxon $p = 9.54 \times 10^{-7}$). Full results appear in Sections E and I.

5.6 Efficiency and Robustness

Efficiency. PURIST combines low training time with modest memory overhead (Section K). For soft-target storage, SAEM needs only one scalar per unlabeled example, requiring $O(|\mathcal{U}|)$ auxiliary memory (Section J).

Robustness. PURIST varies little across random seeds: its mean per-dataset AUC standard deviation is among the three lowest for PU methods in each SAR regime (Section D). Across five labeling rates, it remains within 1.0 AUC percentage point of the best compared PU method and achieves the highest macro-average AUC at c=0.05 and c=0.1 (Table 5). Sensitivity to $\gamma_{\min}$ is limited over the tested range; full analyses appear in Section H.

Table 5: Label-rate sensitivity (SCAR). AUC is averaged over five paired seeds within each dataset, then equally across 10 datasets. †: PN oracle; bold: best PU result.

Method	c=0.01	c=0.05	c=0.1	c=0.3	c=0.5
PN†	97.8	97.7	97.8	98.1	98.2
PURIST	83.0	**93.1**	**95.3**	96.2	96.5
DT	79.2	90.9	94.0	**97.0**	**97.5**
nnPUSB	**83.6**	91.6	92.7	96.3	96.9
VPU	81.6	89.6	92.6	96.1	96.5
LBE	66.1	86.4	91.7	94.5	95.3

6 Discussion and Conclusion

Design implications. PU learning provides asymmetric supervision: labeled positives are trusted, whereas unlabeled examples contain hidden positives. PURIST reflects this structure by balancing the two sides during optimization and refining only unlabeled targets. Under SCAR, population-optimal P–U BCE preserves the Bayes ranking at any fixed sampling ratio, allowing sampling to address finite-sample optimization while retaining this population property. These observations motivate jointly considering supervision reliability, sampling, and target refinement.

When temporal refinement helps. Dai et al. (2025) report that moving-average pseudo-labeling often reduces accuracy in their evaluated configurations, attributing this behavior to unreliable historical predictions. Our controlled SCAR ablation adds evidence that sampling conditions matter: the same SAEM schedule yields negligible average AUC gain under natural P:U proportions but a significant gain under balanced sampling, with a significant positive interaction between the two components. The early target diagnostic also shows greater reductions in gradient-coefficient distortion under balanced sampling. Together, these findings support evaluating temporal refinement jointly with the training conditions that produce its prediction history.

Beyond SCAR. We extend the analysis to instance-dependent positive selection. Under SAR, population-optimal P–U BCE ranks examples by the propensity–posterior product $e(x)\eta(x)$. Our analysis identifies conditions for Bayes-optimal ranking, bounds AUC regret under selection bias, and establishes that zero propensity–posterior correlation is insufficient for identifiability. Under exact population updates, finite-time SAEM preserves the initial ranking and therefore inherits its guarantees and limitations. This analysis explains how labeling mechanisms affect the method's ranking properties and complements the experiments under three controlled SAR constructions.

Limitations. PURIST uses a shared design across tasks, and task-specific architectures or augmentation may further improve individual applications. Our SAR experiments cover three proxy-based constructions; performance under other selection mechanisms requires further study. Our primary focus is AUC-based ranking; applications requiring calibrated probabilities or threshold-based decisions may additionally need calibration or threshold selection.

Conclusion. PURIST combines a simple design with strong performance, empirical robustness, and low computational overhead. Using a single model without a supplied class prior or data augmentation, it achieves the best average rank across 21 SCAR datasets and three SAR constructions on 14 datasets. Its stability across random seeds and labeling rates, together with low training cost and modest auxiliary memory, supports its practical value for heterogeneous PU tasks. These findings demonstrate the effectiveness of jointly designing balanced optimization and U-only target refinement, while our SAR analysis clarifies the ranking guarantees and their boundaries under selection bias.

AI USE STATEMENT

Generative AI tools assisted with refining the conceptual framing and hypotheses; providing feedback on methodology and experimental design; formulating and checking mathematical statements and proof drafts; implementing and reviewing research code; interpreting experimental results; translating and editing manuscript text; and organizing tables and references. They were also used for literature discovery and summarization. The authors manually reviewed all AI-assisted text and citations, checked the mathematical arguments, and verified code and numerical results against the experiment outputs. The authors take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Method definitions, assumptions, and theoretical results are presented in Sections 3 and 4, with complete proofs in Section A and the training algorithm, optimization settings, and random-seed protocol in Section F. Dataset sources, task definitions, preprocessing, and backbones are described in Section C, and the SAR constructions in Section B. The experimental protocol, baseline configurations and adaptations, and checkpoint-selection policy are documented in Sections 5.1 and G. Full results, ranking conventions, and the F1 decision-threshold protocol appear in Section D; metric scales and aggregation rules are specified in Section 3.1 and the corresponding table captions. Component ablations and their statistical procedures are given in Section E, sensitivity analyses in Section H, training diagnostics in Section I, and memory and training-time comparisons in Sections J and K. These descriptions can also guide AI agents in reimplementing the method, reproducing experiments, and checking outputs against the reported results.

REFERENCES

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2019)*, volume 11907 of *Lecture Notes in Computer Science*, pp. 71–85. Springer, 2020. doi: 10.1007/978-3-030-46147-8_5.

Dainis A. Boumber, Fatima Zahra Qachfar, and Rakesh Verma. Domain-agnostic adapter architecture for deception detection: Extensive evaluations with the DIFrauD benchmark. In *LREC-COLING*, pp. 5260–5274, 2024. URL https://aclanthology.org/2024.lrec-main.468/.

Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *NeurIPS*, 2020a.

Qiuyi Chen, Haiyang Zhang, Leqi Zhang, Changchun Li, Jia Wang, and Wei Wang. PU-Bench: A unified benchmark for rigorous and reproducible PU learning. In *ICLR*, 2026. URL https://openreview.net/forum?id=tb8DabMbMq.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In *ICML*, 2020b.

Songmin Dai, Xiaoqiang Li, Yue Zhou, Xichen Ye, and Tong Liu. GradPU: Positive-unlabeled learning via gradient penalty and positive upweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7296–7303, 2023.

Yuanchao Dai, Zhengzhang Hou, Changchun Li, Yuanbo Xu, En Wang, and Ximing Li. A closer look to positive-unlabeled learning from fine-grained perspectives: An empirical study. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5672.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *NeurIPS*, 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.

Emanuele Francazi, Marco Baity-Jesi, and Aurelien Lucchi. A theoretical analysis of the learning dynamics under class imbalance. In *ICML*, volume 202, pp. 10285–10322. PMLR, 2023.

Konrad Furmańczyk, Jan Mielniczuk, Wojciech Rejchel, and Paweł Teisseyre. Double logistic regression approach to biased positive-unlabeled data. In *ECAI*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 764–771, 2023. doi: 10.3233/FAIA230342.

Saurabh Garg, Yifan Wu, Alexander Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture proportion estimation and PU learning: A modern approach. In *NeurIPS*, 2021.

Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke Rundensteiner. Recovering the propensity score from biased positive unlabeled data. In *AAAI*, pp. 6694–6702, 2022. doi: 10.1609/aaai.v36i6.20624.

Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane You, Jian Yang, and Dacheng Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4163–4177, 2022. doi: 10.1109/TPAMI.2021.3061456.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. Predictive adversarial learning from positive and unlabeled data. In *AAAI*, volume 35, pp. 7806–7814, 2021.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: Beyond empirical risk minimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/e0ab531ec312161511493b002f9be2ee-Abstract.html.

Alon Jacovi, Gang Niu, Yoav Goldberg, and Masashi Sugiyama. Scalable evaluation and improvement of document set expansion via neural positive-unlabeled learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 581–592. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.47. URL https://aclanthology.org/2021.eacl-main.47/.

Yangbangyan Jiang, Qianqian Xu, Yunrui Zhao, Zhiyong Yang, Peisong Wen, Xiaochun Cao, and Qingming Huang. Positive-unlabeled learning with label distribution alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15345–15363, 2023.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *ICLR*, 2019.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2023. URL https://archive.ics.uci.edu/.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

Ken Lang. NewsWeeder: Learning to filter netnews. In *ICML*, pp. 331–339, 1995.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Changchun Li, Ximing Li, Lei Feng, and Jihong Ouyang. Who is your right mixup partner in positive and unlabeled learning. In *ICLR*, 2022. URL https://openreview.net/pdf/d687929225415551919e783b0c52f61382054101.pdf.

Changchun Li, Yuanchao Dai, Lei Feng, Ximing Li, Bing Wang, and Jihong Ouyang. Positive and unlabeled learning with controlled probability boundary fence. In *ICML*, volume 235, pp. 27641–27652, 2024.

Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 2391–2404, 2023. doi: 10.18653/v1/2023.acl-long.134.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, pp. 179–186, 2003. doi: 10.1109/ICDM.2003.1250918.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 2020.

Lin Long, Haobo Wang, Zhijie Jiang, Lei Feng, Chang Yao, Gang Chen, and Junbo Zhao. Positive-unlabeled learning by latent group-aware meta disambiguation. In *CVPR*, pp. 23138–23147, 2024.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL-HLT*, pp. 142–150, 2011. URL https://aclanthology.org/P11-1015/.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. MLlib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17(34):1–7, 2016. URL https://www.jmlr.org/papers/v17/15-237.html.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, 2015.

Byeonghu Na, Hyemi Kim, Kyungwoo Song, Weonyoung Joo, Yoon-Yeong Kim, and Il-Chul Moon. Deep generative positive-unlabeled learning under selection bias. In *CIKM*, 2020.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. Efficient large-scale language model training on GPU clusters using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. Association for Computing Machinery, 2021. doi: 10.1145/3458817.3476209.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Carlos Ortega Vázquez, Jochen De Weerdt, and Seppe vanden Broucke. The hidden cost of fraud: An instance-dependent cost-sensitive approach for positive and unlabeled learning. In *Proceedings of the Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 183 of *Proceedings of Machine Learning Research*, pp. 53–67. PMLR, 2022. URL https://proceedings.mlr.press/v183/vasquez22a.html.

Aleš Papič, Igor Kononenko, and Zoran Bosnić. Conditional generative positive and unlabeled learning. *Expert Systems with Applications*, 224:120046, 2023. doi: 10.1016/j.eswa.2023.120046.

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *KDD*, pp. 985–994, 2015. doi: 10.1145/2783258.2783370.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *ICISSP*, pp. 108–116, 2018. doi: 10.5220/0006639801080116.

Keyu Song, Chao Liu, and Dongxiang Jiang. A positive-unlabeled learning approach for industrial anomaly detection based on self-adaptive training. *Neurocomputing*, 647:130488, 2025. doi: 10.1016/j.neucom.2025.130488.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

Xinrui Wang, Wenhai Wan, Chuanxing Geng, Shao-Yuan Li, and Songcan Chen. Beyond myopia: Learning from positive and unlabeled data through holistic predictive trends. In *NeurIPS*, volume 36, pp. 67589–67602, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d5c0f9585592bad5251133813893a6c0-Abstract-Conference.html.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. doi: 10.48550/arXiv.1804.03209.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019. doi: 10.48550/arXiv.1908.02591.

Jonathan Wilton, Abigail M. Y. Koay, Ryan K. L. Ko, Miao Xu, and Nan Ye. Positive-unlabeled learning using random forests via recursive greedy risk minimization. In *NeurIPS*, pp. 24060–24071, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

Zheng Xie, Yu Liu, Hao-Yuan He, Ming Li, and Zhi-Hua Zhou. Weakly supervised AUC optimization: A unified partial AUC approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4780–4795, 2024. doi: 10.1109/TPAMI.2024.3357814.

William J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1⟨32::AID-CNCR2820030106⟩3.0.CO;2-3.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, pp. 649–657, 2015.

Hengwei Zhao, Zhengzhong Tu, Zhuo Zheng, Wei Wang, Junjue Wang, Rusty Feagin, and Wenzhe Jiao. Noisy-pair robust representation alignment for positive-unlabeled learning. In *ICLR*, 2026.

Yunrui Zhao, Qianqian Xu, Yangbangyan Jiang, Peisong Wen, and Qingming Huang. Dist-PU: Positive-unlabeled learning from a label distribution perspective. In *CVPR*, 2022.

Zhangchi Zhu, Lu Wang, Pu Zhao, Chao Du, Wei Zhang, Hang Dong, Bo Qiao, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Robust positive-unlabeled learning via noise negative sample self-correction. In *KDD*, 2023. doi: 10.1145/3580305.3599491.

A PROOFS FOR THE POPULATION-RANKING RESULTS

Proposition (Fixed-ratio P–U BCE ranking under SAR)**.** *Fix any P sampling prior $\alpha \in (0,1)$ and assume $e(x) > 0$ for p_+-almost every x. Under SAR-A, the population-optimal BCE probability $w^*_{e,\alpha}$ achieves $\mathrm{AUC} = \mathrm{AUC}^*$ without estimating propensity. Under SAR-B, it ranks by $e(x)\eta(x)$ and satisfies $\mathrm{AUC}^* - \mathrm{AUC}(w^*_{e,\alpha}) \le \min\{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)/(1-\pi), 2\,\mathrm{AUC}^* - 1\}$. Under SAR-C (orthogonal propensity with both variables of positive variance, $\mathrm{Corr}_{X\sim q}(e(X), \eta(X)) = 0$), the observed distribution identifies only the normalized product $e\eta$; the two factors remain non-identifiable in general.*

We prove each case in turn. Throughout, all densities are taken with respect to a common dominating measure; we write $p_{\mathcal{P}}(x)$ for the density of observed (labeled) positives and $p_+(x)$ for the true positive-class density. Under SAR the two are related by

$$p_{\mathcal{P}}(x) \;=\; \frac{e(x)\,p_+(x)}{\mathbb{E}_{X\sim p_+}[e(X)]}, \tag{3}$$

where $e(x) = P(S{=}1 \mid Y{=}1, X{=}x)$ is the instance-dependent labeling propensity.

Proof. **Case 1: SAR-A (monotone propensity).** Assume $\eta(x_1) \ge \eta(x_2)$ implies $e(x_1) \ge e(x_2)$, following the SAR-A regime of Gong et al. (2022).

Step 1 (density ratio preserves strict posterior order). Fixed-ratio BCE treats $\mathcal{P}$ vs. $\mathcal{U}$ as the two classes with sampling prior α. Its population-optimal probability output is

$$w^*_{e,\alpha}(x) = \frac{\alpha p_{\mathcal{P}}(x)}{\alpha p_{\mathcal{P}}(x) + (1-\alpha) q(x)},$$

which is strictly increasing in the density ratio $\rho(x) \coloneqq p_{\mathcal{P}}(x)/q(x)$. Using $p_+(x) = \eta(x)q(x)/\pi$ and (3), q-almost everywhere,

$$\rho(x) \;=\; \frac{e(x)\eta(x)}{\pi\,\mathbb{E}_{X\sim p_+}[e(X)]} \;=\; \frac{e(x)\eta(x)}{\mathbb{E}_{X\sim q}[e(X)\eta(X)]}.$$

If $\eta(x_1) > \eta(x_2) > 0$, then SAR-A and positive propensity give $e(x_1)\eta(x_1) \ge e(x_2)\eta(x_1) > e(x_2)\eta(x_2)$. If $\eta(x_2) = 0$, the product at x_2 is zero and the product at x_1 is positive. Thus ρ strictly preserves every strict posterior ordering, up to null sets.

Step 2 (AUC optimality). Since $w^*_{e,\alpha}$ is strictly increasing in ρ, it strictly preserves every strict ordering of η. Such a scorer achieves the Bayes-optimal AUC: $\mathrm{AUC}(w^*_{e,\alpha}) = \mathrm{AUC}(\eta) = \mathrm{AUC}^*$.

Remark (SCAR as special case). SCAR sets $e(x) = c$ for all x, which is trivially monotone. Then $p_{\mathcal{P}} = p_+$, $\rho(x) \propto p_+(x)/q(x)$, and the result is a case-control instance of the AUC corruption-invariance result of Menon et al. (2015).

Case 2: SAR-B (anti-monotone propensity). Following the SAR-B regime of Gong et al. (2022), propensity decreases with the posterior; the argument below in fact applies to any positive propensity. Since

$$\frac{p_{\mathcal{P}}(x)}{q(x)} = \frac{e(x)\eta(x)}{\mathbb{E}_{X\sim q}[e(X)\eta(X)]},$$

the population-optimal BCE probability

$$w^*_{e,\alpha}(x) \;=\; \frac{\alpha p_{\mathcal{P}}(x)}{\alpha p_{\mathcal{P}}(x) + (1-\alpha) q(x)},$$

is strictly increasing in $e(x)\eta(x)$ and maximizes $\mathrm{AUC}(f; p_{\mathcal{P}}, q)$. Let

$$\Delta \;\coloneqq\; \mathrm{AUC}^* - \mathrm{AUC}(w^*_{e,\alpha};\, p_+, r), \tag{4}$$

and let f^*_0 be any scorer strictly increasing in η, which is optimal for both (p_+, q) and (p_+, r).

For fixed f and reference distribution R, the AUC integrand lies in $[0,1]$, so

$$|\,\mathrm{AUC}(f; P_1, R) - \mathrm{AUC}(f; P_2, R)| \le \mathrm{TV}(P_1, P_2). \tag{5}$$

Let $\delta_e = \mathrm{TV}(p_{\mathcal{P}}, p_+)$. Since $q = \pi p_+ + (1-\pi)r$ and the AUC of a distribution against itself is $1/2$,

$$\mathrm{AUC}(f; p_+, q) = \frac{\pi}{2} + (1-\pi)\,\mathrm{AUC}(f; p_+, r). \tag{6}$$

The scorer f_0^* ranks by $p_+/q = \eta/\pi$ and hence achieves AUC^* on (p_+, r). Using the optimality of $w_{e,\alpha}^*$ for $(p_{\mathcal{P}}, q)$ and applying (5) twice,

$$\begin{aligned}
&(1-\pi)\Delta \\
&= \mathrm{AUC}(f_0^*; p_+, q) - \mathrm{AUC}(w_{e,\alpha}^*; p_+, q) \\
&\le \left|\mathrm{AUC}(f_0^*; p_+, q) - \mathrm{AUC}(f_0^*; p_{\mathcal{P}}, q)\right| + \left|\mathrm{AUC}(w_{e,\alpha}^*; p_{\mathcal{P}}, q) - \mathrm{AUC}(w_{e,\alpha}^*; p_+, q)\right| \\
&\le 2\delta_e.
\end{aligned} \tag{7}$$

Therefore

$$\Delta \le \frac{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)}{1-\pi}. \tag{8}$$

The AUC loss also has an exact pairwise form. Let $(U, T) = (\eta(X), e(X)\eta(X))$ for $X \sim q$, let (U', T') be an independent copy, and define

$$L_T = \mathbf{1}\{(U-U')(T-T') < 0\} + \tfrac{1}{2}\mathbf{1}\{U \neq U',\, T = T'\}.$$

Thus L_T assigns unit cost to a reversed posterior pair and half cost when distinct posterior values receive tied scores. For $H_f(X, X') = \mathbf{1}\{f(X) > f(X')\} + \frac{1}{2}\mathbf{1}\{f(X) = f(X')\}$, symmetrizing the usual AUC expression gives

$$\mathrm{AUC}(\eta) - \mathrm{AUC}(f) = \frac{\mathbb{E}[(U-U')(H_\eta - H_f)]}{2\pi(1-\pi)}.$$

Substituting any scorer ordered by T yields

$$\Delta = \frac{\mathbb{E}[|U-U'|L_T]}{2\pi(1-\pi)}. \tag{9}$$

Setting $f = -\eta$ in the same identity shows that $\mathbb{E}|U-U'|/[2\pi(1-\pi)] = 2\,\mathrm{AUC}^* - 1$. Since $0 \le L_T \le 1$, combining this intrinsic rank-signal cap with (8) gives

$$\Delta \le \min\left\{\frac{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)}{1-\pi},\; 2\,\mathrm{AUC}^* - 1\right\} \le 1. \tag{10}$$

Equation (9) distinguishes selection shift from its ranking effect: only the mass and posterior gaps of reversed or tied pairs contribute to AUC degradation.

Case 3: SAR-C (orthogonal propensity). Assume both factors have positive variance and $\mathrm{Corr}_{X\sim q}(e(X), \eta(X)) = 0$.

Step 1 (the case-control observation model). Because $p_+(x) = \eta(x)q(x)/\pi$, substituting into (3) gives

$$p_{\mathcal{P}}(x) = \frac{e(x)\eta(x)q(x)}{\mathbb{E}_{X\sim q}[e(X)\eta(X)]}. \tag{11}$$

Thus, given q, the observed-positive distribution identifies only the normalized product of e and η.

Step 2 (an admissible SAR-C counterexample with incompatible rankings). Let $X = (A, B)$, where A and B are independent $\mathrm{Bernoulli}(1/2)$ variables under q, and define

$$e(X) = 0.4 + 0.2A, \qquad \eta(X) = 0.2 + 0.1B.$$

Both variables have positive variance, and independence of A and B implies $\mathrm{Corr}_{X\sim q}(e(X), \eta(X)) = 0$. Now define a second pair

$$\tilde{e}(X) = 2\eta(X) = 0.4 + 0.2B, \qquad \tilde{\eta}(X) = \frac{e(X)}{2} = 0.2 + 0.1A.$$

Both pairs have propensities in $\{0.4, 0.6\}$ and posteriors in $\{0.2, 0.3\}$. Hence both are admissible, have positive variances, satisfy the zero-correlation constraint, and obey

$$\tilde{e}(x)\tilde{\eta}(x) = e(x)\eta(x) \quad \text{for every } x.$$

By (11), they induce the same $(p_{\mathcal{P}}, q)$ and have the same class prior $\mathbb{E}_q[\eta(X)] = \mathbb{E}_q[\tilde{\eta}(X)] = 0.25$. However, at $x = (0, 1)$ and $x' = (1, 0)$, $\eta(x) = 0.3 > 0.2 = \eta(x')$, whereas $\tilde{\eta}(x) = 0.2 < 0.3 = \tilde{\eta}(x')$. Both points have positive probability under q, so the two observationally indistinguishable models require incompatible Bayes rankings. Thus neither the factor pair nor the Bayes ranking is generally identifiable under SAR-C, even if the class prior is known.

Summary. The three cases yield the following conclusions:

1. **SAR-A:** when e is monotone in η, the population-optimal probability output of fixed-ratio P–U BCE achieves $\mathrm{AUC} = \mathrm{AUC}^*$ without estimating propensity.
2. **SAR-B:** the ordering need not be preserved; the bound $\min\{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)/(1-\pi), 2\,\mathrm{AUC}^* - 1\}$ applies. Whether an anti-monotone propensity is harmful depends on whether $e\eta$ reverses posterior pairs.
3. **SAR-C:** zero correlation does not separate propensity from posterior because only their normalized product is observed; even a known class prior does not generally identify the Bayes ranking.

□

A.1 Finite-Time SAEM Rank Preservation

For a current unlabeled target $\phi_t(x)$, minimize pointwise population balanced BCE over measurable probability scores in $[0, 1]$, with the usual boundary convention for BCE. On the support where $p_{\mathcal{P}}(x) + q(x) > 0$, the optimum is

$$q_t(x) = \frac{p_{\mathcal{P}}(x) + q(x)\phi_t(x)}{p_{\mathcal{P}}(x) + q(x)} = w(x) + \big(1 - w(x)\big)\phi_t(x). \tag{12}$$

Substitution into the target refresh gives

$$\phi_{t+1}(x) = \phi_t(x) + \gamma_{t+1} w(x)\big(1 - \phi_t(x)\big), \tag{13}$$

and hence

$$1 - \phi_{t+1}(x) = \big(1 - \gamma_{t+1} w(x)\big)\big(1 - \phi_t(x)\big). \tag{14}$$

Starting from $\phi_0 = 0$ and iterating this identity yields

$$\phi_t(x) = 1 - \prod_{s=1}^{t} \big(1 - \gamma_s w(x)\big).$$

Equation (12) then gives the stated formula for q_t:

$$q_t(x) = 1 - \big(1 - w(x)\big) \prod_{s=1}^{t} \big(1 - \gamma_s w(x)\big).$$

For $0 < \gamma_s \le 1$ and $0 \le w_1 < w_2 \le 1$, every factor at w_1 is strictly larger than its counterpart at w_2 and remains positive; if a factor at w_2 reaches zero, the product comparison is still strict. Thus the product is strictly decreasing and ϕ_t is strictly increasing for $t \ge 1$. Treating $(1 - w)$ as an additional factor gives the same strict comparison for q_t for every finite t, including $t = 0$. Hence both the strict order and ties induced by w are preserved.

Finally, if $\sum_s \gamma_s = \infty$ and $w(x) > 0$, then

$$\prod_{s=1}^{t} \big(1 - \gamma_s w(x)\big) \le \exp\left(-w(x) \sum_{s=1}^{t} \gamma_s\right) \longrightarrow 0.$$

Thus both $\phi_t(x)$ and $q_t(x)$ approach one on that support, establishing the limiting score compression stated in Section 4.2.

B SAR CONSTRUCTIONS AND RANKING IMPLICATIONS

The fixed-ratio P–U BCE population-ranking result (Proposition 4.1) distinguishes three labeling-bias scenarios.

SAR-A (monotone propensity). When $e(x) > 0$ p_+-almost surely and is nondecreasing in $\eta(x)$, the observed-positive-to-unlabeled density ratio $p_{\mathcal{P}}(x)/q(x) \propto e(x)\eta(x)$ strictly preserves the strict ordering of $\eta(x)$. The population P–U BCE scorer achieves $\mathrm{AUC} = \mathrm{AUC}^*$ without knowing e. This regime follows the increasing-propensity construction of Gong et al. (2022).

SAR-B (anti-monotone propensity). When $e(x)$ decreases in $\eta(x)$, the ordering is preserved exactly on pairs for which the product $e(x)\eta(x)$ remains in posterior order. Anti-monotonicity alone therefore permits both order-preserving and fully reversed cases. In general, the AUC degradation is bounded by $\min\{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)/(1-\pi), 2\,\mathrm{AUC}^* - 1\}$, whose terms measure selection-induced shift and intrinsic rank signal. Meanwhile, (9) counts the posterior-gap mass of pairs whose $e\eta$ order is reversed or tied. The decreasing-propensity regime follows Gong et al. (2022).

SAR-C (orthogonal propensity, $\mathrm{Corr}_{X\sim q}(e(X), \eta(X)) = 0$). Only the normalized product $e\eta$ is observed, so zero correlation does not make the two factors separately identifiable (Section A).

Experimental construction. We retain PU-Bench's auxiliary logistic-regression proxy: scikit-learn `LogisticRegression` with the L-BFGS solver, at most 100 iterations, L2 regularization, and $C = 1$. It is fitted to the fully labeled source features (flattened when necessary) and evaluated on that same source split. Training and validation proxies are fitted separately, before PU sampling; test data are not used. This proxy defines the synthetic labeling mechanism and is distinct from the neural PN oracle in the result tables. The inherited implementation is pinned in Section C; SAR-C adds the residualization and dependent-rounding construction below. Let $s_i \in [0, 1]$ be this proxy for positive example i, and let n_+ be the number of positives in the source pool. Observed positives are drawn as a fixed-size weighted sample without replacement, using normalized selection weights

$$q_A(i) = 0.9\frac{s_i^{10}}{\sum_j s_j^{10}} + \frac{0.1}{n_+}, \qquad q_B(i) = \frac{(2.5 - s_i)^{10}}{\sum_j (2.5 - s_j)^{10}}. \tag{15}$$

Because sampling is without replacement, q_A and q_B are selection weights rather than observed-positive marginals. If $\alpha_R(i)$ is the first-order inclusion probability under regime $R \in \{A, B\}$ and m points are drawn, the induced finite-pool observed-positive distribution assigns mass $\alpha_R(i)/m$ to point i. The weights q_A favor high proxy scores, whereas q_B favors low proxy scores. At the weight level, the SAR-B product $s(2.5 - s)^{10}$ has derivative with the sign of $2.5 - 11s$ and therefore reverses direction over $s > 2.5/11$.

For SAR-C, let $a_i = -\|x_i - \mu_+\|_2$ be proximity to the source positive centroid and let bars denote source-population means. We remove its linear component along the PN proxy,

$$\widetilde{a}_i = (a_i - \bar{a}) - \frac{\sum_j (a_j - \bar{a})(s_j - \bar{s})}{\sum_j (s_j - \bar{s})^2}(s_i - \bar{s}), \qquad r_i = \frac{\widetilde{a}_i}{\max_j |\widetilde{a}_j|}.$$

The prescribed propensity is $e_C(i) = \kappa(1 + 0.2 r_i)$, where κ is chosen so that the positive-pool propensities sum to the requested number of labeled positives. Dependent rounding then draws a fixed-size sample without replacement with these first-order inclusion probabilities. Affine scaling preserves the residualization, so e_C has zero empirical covariance with s over the source population; this defines the proxy-orthogonal SAR-C experiment.

B.1 THEORY–EXPERIMENT ALIGNMENT DIAGNOSTICS

The SAR-A/B relations to s are exact construction properties: q_A increases with s, while the SAR-B product $s(2.5 - s)^{10}$ decreases above $2.5/11$. On real source data, the pointwise posterior $\eta(x)$ remains latent when the class label Y is observed. We therefore evaluate the construction proxy against the source labels using AUC and Brier score. For SAR-C, the population identity

$$\mathrm{Cov}(e(X), Y) = \mathrm{Cov}(e(X), \mathbb{E}[Y \mid X]) = \mathrm{Cov}(e(X), \eta(X))$$

Table 6: Alignment diagnostics for the proxy-defined SAR constructions, averaged over ten seeds. s is the PN proxy used to construct SAR-A/B/C. Proxy AUC and Brier score are measured against the fully observed source labels. SAR-C is exactly orthogonal to s in empirical covariance; the final column reports its absolute source-label correlation. Since $\mathbb{E}[Y \mid X] = \eta(X)$, $\mathrm{Cov}(e(X), Y) = \mathrm{Cov}(e(X), \eta(X))$ in the population.

Dataset	$\mathrm{AUC}(s, Y)$	$\mathrm{Brier}(s, Y)$	$\lvert\widehat{\mathrm{Corr}}(e_C, Y)\rvert$
Adult	0.907	0.102	0.029
CIFAR10	0.894	0.125	0.049
Connect4	0.863	0.142	0.000
DBpedia	0.994	0.013	0.029
DiFraud	0.838	0.159	0.002
Elliptic	0.967	0.037	0.014
FashionMNIST	0.997	0.014	0.011
MNIST	0.962	0.073	0.015
Mushrooms	1.000	0.000	0.011
Pendigits	0.944	0.086	0.004
SUSY	0.858	0.148	0.044
Spambase	0.977	0.055	0.030
SpeechCommands	0.739	0.134	0.018
YelpChi	0.773	0.105	0.010
Mean	0.908	0.085	0.019

turns the observed source labels into a direct diagnostic of the theoretical zero-covariance moment. Table 6 reports these quantities across the real-data panel.

Across the 14 dataset means, the source-label AUC of s is 0.908. The absolute SAR-C source-label correlation averages 0.019 and reaches at most 0.049.

For an exact finite-support illustration, let $U \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and $Z \in \{-1, 1\}$ be independent and uniform, set $X = (U, Z)$ and $\eta(X) = U$, and use $e_A = 0.2 + 0.6U$, $e_B = 0.8(2.5 - U)^{10}/2.4^{10}$, and $e_C = 0.5 + 0.3Z$. Table 7 evaluates the monotone, anti-monotone, and zero-correlation conditions by finite summation, together with exact pairwise rank corruption and the bound in Proposition 4.1.

Table 7: Exact population illustration on a finite-support distribution with known posterior η. $\Delta_e = \mathrm{AUC}^* - \mathrm{AUC}(e\eta)$, and $\kappa_e = \Delta_e/(2\,\mathrm{AUC}^* - 1)$ is the posterior-gap-weighted rank-corruption rate in (9), with $\kappa_e = 0$ when $\mathrm{AUC}^* = 1/2$. $B_e = \min\{2\,\mathrm{TV}(p_{\mathcal{P}}, p_+)/(1-\pi), 2\,\mathrm{AUC}^* - 1\}$.

Regime	$\mathrm{Corr}(e, \eta)$	TV	κ_e	Δ_e	B_e
SAR-A	+1.000	0.100	0.000	0.000	0.402
SAR-B	-0.924	0.528	0.964	0.648	0.672
SAR-C	+0.000	0.300	0.140	0.094	0.672

C Dataset Details

Benchmark inheritance and data splits. For datasets already supported by PU-Bench (Chen et al., 2026), we retain its loaders, binary class mappings, and source train/test splits, pinned to revision `2d95a19eefd7`, with the adaptations specified here and in Section G. Table 8 gives the expanded panel and its positive classes. The added CIFAR100 and SVHN loaders use the official train/test splits and pixel values divided by 255. Adult, Covertype, SUSY, Pendigits, Letter, Elliptic, YelpChi, and CICIDS2017 use stratified 80/20 source train/test splits with the experiment seed. DBpedia combines the released train/test pools, subsamples 50K examples stratified by the original 14 classes, and then uses a binary-stratified 80/20 split; its text concatenates title and content. DiFraud combines the released splits and uses a stratified 80/20 split. These two text caches are constructed with seed 42. From each training source pool, 1% is held out for validation by stratified sampling with the experiment seed, before constructing P and U separately within the training and validation splits. Cached source train/test splits are retained across seeds; the validation and PU sampling use the listed experiment seeds. SUSY uses a stratified 500K subset of the original 5M rows. For SpeechCommands, the evaluated cache uses the 9,981 utterances in the official validation list as the held-out test set, and the remaining 95,848 utterances as the source training pool, from which the 1% validation split is drawn.

P/U overlap. We retain the upstream full-U case-control convention (`nnpu_full_u`, also named `naive_mode`). In each source split, $\mathcal{P}$ contains $\lfloor cn_+ \rfloor$ selected positive examples, while $\mathcal{U}$ contains every source example with its label hidden. Thus each selected positive also has an unlabeled copy in $\mathcal{U}$; P and U are not disjoint source subsets. Selection within P is without replacement, uniform under SCAR and as specified in Section B under SAR. Source identities remain disjoint between training, validation, and test.

Shared backbones. We retain the upstream 4-layer multilayer perceptron with dropout for tabular and text data, LeNet for MNIST/FashionMNIST, and CNN-CIFAR10 for CIFAR-10/100/SVHN, including their layer widths and activations. Text inputs retain the frozen 384-dimensional embeddings from `sentence-transformers/all-MiniLM-L6-v2`; only the downstream classifier is trained in the PU experiments. DBpedia and DiFraud use the same encoder and per-example L2 normalization. Cached embeddings and their stored labels are loaded together to preserve row alignment. SpeechCommands uses 40×101 log-Mel spectrograms and a two-layer time–frequency CNN trained end to end. Audio is resampled to 16 kHz and padded or truncated to one second; the spectrogram uses 40 Mel bands, a 400-sample window, and a 160-sample hop. Both convolutional layers have 64 channels, ReLU activation, and dropout 0.5. Their frequency–time kernels are 8×20 and 4×10, respectively, with stride one and same padding. A 2×2 max pool with ceiling rounding follows the first layer; the final $64 \times 20 \times 51$ tensor is flattened into a single-logit linear head. The same representation is used in SCAR, SAR, ablations, and dynamics, with method-specific heads or mixing adapters where required. The power spectrogram is converted to decibels with an 80 dB dynamic range and standardized using a scalar mean and standard deviation from the source training pool. For Elliptic and YelpChi, we retain the source split identities but fit median imputation and standardization only on the training source split before PU sampling, excluding validation and test rows. We use a two-layer mean GraphSAGE encoder (Hamilton et al., 2017) with 64 hidden units per layer, ReLU activations, and a linear output head. Uniform neighbor sampling with replacement uses fanouts 10 and 5; isolated nodes have zero neighbor contributions. The protocol is transductive: all node features and edges provide context, but optimization losses use training nodes except for LaGAM's inherited validation-support loss (Section G). Test labels are reserved for evaluation. Elliptic uses symmetrized transaction edges and all 203,769 context nodes, with supervised splits drawn from the 46,564 known-label nodes. We use the processed YelpChi release available from CARE-GNN and DGFraud, containing 45,954 nodes and 32 features, and take the undirected union of its three relations. Self edges are removed. Training resamples neighbors, while evaluation uses fixed seed-specific samples. Method-specific input adapters and the PUET exception are specified in Section G.

Table 8: Dataset details. N is the preprocessed source size before PU sampling; π is the empirical positive ratio in the unlabeled training set. SUSY uses a stratified 500K subset of the original 5M examples. Text datasets use 384-D all-MiniLM sentence embeddings (Reimers & Gurevych, 2019). Elliptic and YelpChi use node features and graph edges, with the PUET exception in Section G. PN AUC is the supervised oracle average on the $100\times$ scale.

Dataset	Source type	N	Dim	π	Positive class	Features	PN AUC
MNIST (LeCun et al., 1998)	Image	70K	784	0.492	even digit	pixel	99.8
FashionMNIST (Xiao et al., 2017)	Image	70K	784	0.500	upper-body	pixel	99.9
CIFAR10 (Krizhevsky, 2009)	Image	60K	3×32×32	0.400	vehicle	pixel	98.5
CIFAR100 (Krizhevsky, 2009)	Image	60K	3×32×32	0.050	vehicles_1	pixel	95.1
SVHN (Netzer et al., 2011)	Image	99.3K	3×32×32	0.539	odd digit	pixel	98.7
Adult (Kelly et al., 2023)	Tabular	48.8K	14	0.239	income>50K	tabular	91.2
Covertype (Kelly et al., 2023)	Tabular	581K	54	0.488	type-2	tabular	97.5
SUSY (Kelly et al., 2023)	Tabular	500K	18	0.458	SUSY signal	tabular	87.5
Pendigits (Kelly et al., 2023)	Tabular	11K	16	0.496	odd digit	tabular	99.8
Letter (Kelly et al., 2023)	Tabular	20K	16	0.194	vowel	tabular	98.4
Spambase (Kelly et al., 2023)	Tabular	4.6K	57	0.394	spam	tabular	96.4
Mushrooms (Kelly et al., 2023)	Tabular	8.1K	22	0.482	poisonous	tabular	100.0
Connect4 (Kelly et al., 2023)	Tabular	67.6K	42	0.658	win	tabular	95.4
IMDB (Maas et al., 2011)	Text	50K	384	0.500	positive	MiniLM	88.7
20News (Lang, 1995)	Text	18.8K	384	0.565	alt/comp/misc/rec	MiniLM	96.6
DBpedia (Zhang et al., 2015)	Text	50K	384	0.071	Company	MiniLM	99.7
Elliptic (Weber et al., 2019)	Graph	46.6K	166	0.098	illicit tx	node/agg. feat.	98.4
DiFraud (Boumber et al., 2024)	Text	95.9K	384	0.389	deceptive	MiniLM	89.5
SpeechCommands (Warden, 2018)	Audio	105.8K	40×101	0.184	yes/no/up/down/go	log-Mel	97.9
YelpChi (Rayana & Akoglu, 2015)	Graph	46K	32	0.145	fraud	review feat.	87.3
CICIDS2017 (Sharafaldin et al., 2018)	Network	2.83M	77	0.197	attack	flow feat.	99.7

D Complete Experimental Tables

The SCAR tables cover all 21 datasets, with ten seeds per configuration except five for CIFAR100. Each SAR-A/B/C table covers the same 25 non-oracle methods on all 14 SAR datasets, with ten seeds per configuration; the PN oracle is also reported. Per-dataset ranks exclude the oracle. Mean ranks average these ranks across datasets.

D.1 SCAR Results

Table 9: Complete test AUC ($\times 100$) (Part 1) (SCAR, c=0.1, mean±std). Best non-oracle in **bold**.

Dataset	PN†	**PURIST**	PseudoLabel	nnPUSB	WSAUC	ELR	P3Mix-E	Holistic	Direct-AUC	DT	PUET	Dist-PU	LaGAM
MNIST	$99.8_{\pm 0.1}$	$99.5_{\pm 0.1}$	$99.6_{\pm 0.1}$	$95.2_{\pm 3.2}$	$99.1_{\pm 1.5}$	$99.3_{\pm 0.5}$	$99.0_{\pm 0.8}$	$98.7_{\pm 0.9}$	$98.5_{\pm 1.3}$	$99.1_{\pm 0.7}$	$99.2_{\pm 0.1}$	**99.8**$_{\pm 0.0}$	$97.6_{\pm 1.0}$
FashionMNIST	$99.9_{\pm 0.1}$	$99.6_{\pm 0.0}$	$99.6_{\pm 0.1}$	$98.6_{\pm 0.7}$	$99.4_{\pm 0.6}$	$99.5_{\pm 0.2}$	$99.4_{\pm 0.3}$	$99.5_{\pm 0.3}$	$98.9_{\pm 0.6}$	$99.2_{\pm 0.4}$	**99.7**$_{\pm 0.0}$	$99.4_{\pm 0.3}$	$99.1_{\pm 0.4}$
CIFAR10	$98.5_{\pm 0.6}$	$95.6_{\pm 0.4}$	$96.0_{\pm 1.2}$	$95.7_{\pm 0.5}$	$95.6_{\pm 1.4}$	$94.9_{\pm 1.1}$	$94.5_{\pm 1.0}$	$97.1_{\pm 0.7}$	$94.2_{\pm 1.7}$	$94.0_{\pm 3.0}$	$88.0_{\pm 0.2}$	$96.7_{\pm 0.6}$	**97.6**$_{\pm 1.3}$
CIFAR100	$95.1_{\pm 0.9}{}^{5}$	$79.0_{\pm 6.2}{}^{5}$	$78.9_{\pm 1.4}{}^{5}$	$82.1_{\pm 2.8}{}^{5}$	$86.6_{\pm 1.5}{}^{5}$	$85.2_{\pm 2.6}{}^{5}$	$71.8_{\pm 3.2}{}^{5}$	$78.5_{\pm 18.1}{}^{5}$	$81.7_{\pm 2.6}{}^{5}$	$77.5_{\pm 3.7}{}^{5}$	$70.3_{\pm 1.4}{}^{5}$	$87.0_{\pm 1.6}{}^{5}$	$86.0_{\pm 1.5}{}^{5}$
SVHN	$98.7_{\pm 0.8}$	$89.2_{\pm 13.2}$	$59.6_{\pm 8.3}$	$93.3_{\pm 1.8}$	$95.4_{\pm 0.9}$	$93.8_{\pm 2.0}$	$60.6_{\pm 16.3}$	**97.3**$_{\pm 0.4}$	$93.4_{\pm 0.8}$	$91.9_{\pm 3.3}$	$75.1_{\pm 0.9}$	$50.0_{\pm 0.0}$	$89.8_{\pm 2.6}$
Adult	$91.2_{\pm 0.2}$	$87.7_{\pm 0.6}$	$88.4_{\pm 0.5}$	$89.3_{\pm 1.3}$	$87.0_{\pm 1.0}$	$87.5_{\pm 1.2}$	$88.7_{\pm 1.2}$	$89.1_{\pm 1.5}$	$86.0_{\pm 1.6}$	$86.8_{\pm 2.0}$	$85.9_{\pm 0.5}$	$89.2_{\pm 0.6}$	$88.6_{\pm 0.7}$
Covertype	$97.5_{\pm 0.5}$	$91.3_{\pm 0.5}$	$90.0_{\pm 0.4}$	$92.2_{\pm 0.3}$	$90.7_{\pm 1.9}$	$91.1_{\pm 0.7}$	$89.9_{\pm 1.2}$	$88.6_{\pm 0.5}$	$93.8_{\pm 0.4}$	$92.8_{\pm 1.2}$	**95.1**$_{\pm 0.1}$	$86.3_{\pm 0.3}$	$89.2_{\pm 0.4}$
SUSY	$87.5_{\pm 0.2}$	**87.0**$_{\pm 0.1}$	$86.8_{\pm 0.2}$	$86.8_{\pm 0.1}$	$84.7_{\pm 0.7}$	$86.0_{\pm 0.3}$	$86.7_{\pm 0.2}$	$86.8_{\pm 0.3}$	$86.6_{\pm 0.3}$	$86.9_{\pm 0.1}$	$85.3_{\pm 0.1}$	$85.7_{\pm 0.3}$	$86.4_{\pm 0.2}$
Pendigits	$99.8_{\pm 0.2}$	$99.7_{\pm 0.2}$	$99.2_{\pm 0.3}$	$96.0_{\pm 3.8}$	$97.0_{\pm 3.4}$	$99.4_{\pm 0.5}$	**99.7**$_{\pm 0.1}$	$99.4_{\pm 0.3}$	$99.4_{\pm 0.3}$	$99.6_{\pm 0.1}$	$99.4_{\pm 0.2}$	$99.2_{\pm 0.2}$	$97.8_{\pm 1.9}$
Letter	$98.4_{\pm 2.4}$	$96.0_{\pm 0.7}$	$95.4_{\pm 0.7}$	$90.5_{\pm 8.0}$	$95.7_{\pm 1.8}$	$94.0_{\pm 2.5}$	$95.9_{\pm 1.3}$	$95.7_{\pm 0.9}$	$95.1_{\pm 1.0}$	$95.6_{\pm 0.5}$	$94.8_{\pm 0.8}$	$89.7_{\pm 2.4}$	$89.6_{\pm 3.4}$
Spambase	$96.4_{\pm 1.1}$	$93.4_{\pm 1.2}$	$93.4_{\pm 0.8}$	$82.9_{\pm 9.3}$	$92.0_{\pm 0.9}$	$93.0_{\pm 2.1}$	$87.6_{\pm 3.3}$	$90.3_{\pm 3.0}$	$90.0_{\pm 3.0}$	$89.3_{\pm 3.4}$	**95.9**$_{\pm 0.6}$	$93.5_{\pm 1.4}$	$90.1_{\pm 4.5}$
Mushrooms	$100.0_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.7_{\pm 0.3}$	$99.9_{\pm 0.1}$	$99.7_{\pm 0.4}$	$99.0_{\pm 1.2}$	$99.9_{\pm 0.1}$	$99.6_{\pm 0.5}$	**100.0**$_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.4_{\pm 1.0}$
Connect4	$95.4_{\pm 1.1}$	$87.8_{\pm 0.7}$	$86.8_{\pm 0.8}$	$86.1_{\pm 1.3}$	$83.9_{\pm 5.6}$	$85.8_{\pm 1.6}$	$86.8_{\pm 1.5}$	$83.2_{\pm 5.5}$	$85.5_{\pm 1.3}$	$85.4_{\pm 1.1}$	$82.8_{\pm 0.7}$	$70.5_{\pm 5.2}$	$82.5_{\pm 2.0}$
IMDB	$88.7_{\pm 0.8}$	$83.2_{\pm 1.2}$	$83.5_{\pm 0.9}$	$84.2_{\pm 1.4}$	$80.2_{\pm 2.9}$	$79.8_{\pm 3.3}$	$80.0_{\pm 4.9}$	$83.6_{\pm 3.1}$	$78.2_{\pm 2.8}$	$77.8_{\pm 4.4}$	$75.5_{\pm 0.5}$	$56.7_{\pm 13.4}$	$82.7_{\pm 1.4}$
20News	$96.6_{\pm 0.9}$	$91.3_{\pm 1.1}$	$91.4_{\pm 1.0}$	$92.2_{\pm 0.6}$	$88.8_{\pm 3.7}$	$90.8_{\pm 1.8}$	$85.4_{\pm 3.4}$	$87.9_{\pm 6.9}$	$88.9_{\pm 2.0}$	$89.0_{\pm 1.8}$	$87.9_{\pm 0.6}$	**92.6**$_{\pm 1.1}$	$89.9_{\pm 6.6}$
DBpedia	$99.7_{\pm 0.1}$	$98.7_{\pm 0.2}$	$98.5_{\pm 0.3}$	$98.9_{\pm 0.2}$	$98.8_{\pm 0.2}$	$97.3_{\pm 1.5}$	$98.5_{\pm 0.7}$	$98.3_{\pm 0.7}$	$97.9_{\pm 0.4}$	$98.0_{\pm 0.3}$	$93.1_{\pm 0.3}$	$99.3_{\pm 0.2}$	$99.2_{\pm 0.1}$
Elliptic	$98.4_{\pm 0.6}$	$96.1_{\pm 0.3}$	$96.1_{\pm 0.4}$	$96.0_{\pm 0.7}$	$96.2_{\pm 0.4}$	$95.9_{\pm 0.8}$	$95.4_{\pm 0.7}$	$88.6_{\pm 13.3}$	$95.2_{\pm 0.7}$	$95.4_{\pm 0.7}$	**96.6**$_{\pm 0.5}$	$94.3_{\pm 0.6}$	$94.3_{\pm 0.8}$
DiFraud	$89.5_{\pm 0.5}$	**83.4**$_{\pm 0.7}$	$82.5_{\pm 0.7}$	$82.6_{\pm 0.4}$	$82.2_{\pm 0.6}$	$79.9_{\pm 1.2}$	$81.9_{\pm 1.8}$	$82.4_{\pm 0.8}$	$80.1_{\pm 1.8}$	$80.1_{\pm 2.1}$	$77.2_{\pm 0.3}$	$66.5_{\pm 16.5}$	$79.7_{\pm 0.7}$
SpeechCommands	$97.9_{\pm 0.4}$	$92.6_{\pm 3.2}$	$63.6_{\pm 17.1}$	$93.3_{\pm 1.2}$	**96.3**$_{\pm 0.3}$	$95.1_{\pm 0.8}$	$90.1_{\pm 13.9}$	$53.9_{\pm 11.8}$	$94.5_{\pm 1.0}$	$88.0_{\pm 4.4}$	$75.1_{\pm 1.0}$	$51.7_{\pm 1.2}$	$86.4_{\pm 5.6}$
YelpChi	$87.3_{\pm 3.1}$	$80.1_{\pm 0.5}$	$79.6_{\pm 1.5}$	$78.0_{\pm 5.1}$	$78.9_{\pm 1.5}$	$79.1_{\pm 1.9}$	$78.4_{\pm 1.5}$	$74.7_{\pm 2.5}$	$79.3_{\pm 2.8}$	$76.9_{\pm 3.8}$	**80.6**$_{\pm 0.7}$	$79.2_{\pm 0.6}$	$76.7_{\pm 1.4}$
CICIDS2017	$99.7_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.5_{\pm 0.1}$	$99.4_{\pm 0.0}$	$99.3_{\pm 0.2}$	$99.4_{\pm 0.1}$	$99.1_{\pm 0.2}$	$99.1_{\pm 0.0}$	$99.5_{\pm 0.0}$	$99.5_{\pm 0.0}$	**99.9**$_{\pm 0.0}$	$98.8_{\pm 0.1}$	$99.3_{\pm 0.1}$
Rank	—	5.0	6.9	8.7	9.0	9.0	10.0	10.4	10.5	10.6	10.8	11.0	11.6

Table 10: Complete test AUC (×100) (Part 2) (SCAR, c=0.1, mean±std). Best non-oracle in **bold**.

Dataset	PULDA	P3Mix-C	SAT	VPU	MeanTeacher	PULCPBF	LBE	BBE-PU	nnPU	Self-PU	PAN	VAE-PU	CGenPU
MNIST	98.7$_{\pm 2.2}$	99.2$_{\pm 0.4}$	99.3$_{\pm 0.3}$	99.0$_{\pm 0.8}$	98.5$_{\pm 1.3}$	97.8$_{\pm 1.6}$	98.6$_{\pm 0.8}$	99.5$_{\pm 0.2}$	98.4$_{\pm 2.2}$	99.4$_{\pm 0.3}$	98.6$_{\pm 0.6}$	70.3$_{\pm 11.3}$	85.2$_{\pm 4.2}$
FashionMNIST	99.1$_{\pm 0.6}$	99.3$_{\pm 0.7}$	99.5$_{\pm 0.1}$	99.0$_{\pm 0.9}$	98.4$_{\pm 2.2}$	98.6$_{\pm 1.1}$	98.7$_{\pm 0.8}$	98.7$_{\pm 2.6}$	97.0$_{\pm 4.9}$	99.1$_{\pm 0.5}$	93.0$_{\pm 7.2}$	77.1$_{\pm 9.5}$	81.6$_{\pm 5.7}$
CIFAR10	96.4$_{\pm 0.9}$	94.8$_{\pm 0.8}$	94.9$_{\pm 1.1}$	94.3$_{\pm 2.2}$	95.2$_{\pm 0.8}$	95.1$_{\pm 0.5}$	88.4$_{\pm 2.8}$	84.4$_{\pm 5.6}$	96.1$_{\pm 0.4}$	92.6$_{\pm 6.1}$	68.9$_{\pm 8.4}$	74.8$_{\pm 8.5}$	68.2$_{\pm 3.3}$
CIFAR100	86.0$_{\pm 2.7}$[5]	73.2$_{\pm 3.6}$[5]	81.4$_{\pm 4.4}$[5]	82.0$_{\pm 4.7}$[5]	61.8$_{\pm 16.4}$[5]	**87.7**$_{\pm 2.4}$[5]	63.0$_{\pm 11.1}$[5]	52.7$_{\pm 6.6}$[5]	63.9$_{\pm 6.7}$[5]	55.1$_{\pm 4.8}$[5]	56.7$_{\pm 6.2}$[5]	52.7$_{\pm 1.2}$[5]	53.2$_{\pm 3.5}$[5]
SVHN	53.3$_{\pm 3.4}$	84.6$_{\pm 3.6}$	94.4$_{\pm 1.0}$	88.0$_{\pm 1.2}$	90.1$_{\pm 4.1}$	50.0$_{\pm 1.0}$	80.0$_{\pm 5.7}$	50.5$_{\pm 0.7}$	52.7$_{\pm 8.9}$	50.0$_{\pm 0.8}$	50.3$_{\pm 0.4}$	51.4$_{\pm 2.3}$	50.3$_{\pm 0.6}$
Adult	88.3$_{\pm 1.3}$	87.5$_{\pm 1.6}$	90.1$_{\pm 0.5}$	90.2$_{\pm 0.6}$	87.2$_{\pm 2.5}$	89.3$_{\pm 0.5}$	**90.4**$_{\pm 0.4}$	88.7$_{\pm 0.5}$	88.0$_{\pm 1.8}$	88.8$_{\pm 1.0}$	88.0$_{\pm 0.8}$	79.6$_{\pm 2.2}$	74.6$_{\pm 5.4}$
Covertype	81.3$_{\pm 1.2}$	87.9$_{\pm 1.1}$	85.5$_{\pm 0.3}$	83.3$_{\pm 0.2}$	91.2$_{\pm 0.3}$	81.8$_{\pm 0.8}$	85.0$_{\pm 0.4}$	78.1$_{\pm 1.2}$	80.2$_{\pm 0.6}$	83.0$_{\pm 0.5}$	81.8$_{\pm 1.0}$	69.4$_{\pm 7.1}$	56.6$_{\pm 3.8}$
SUSY	83.0$_{\pm 0.7}$	86.3$_{\pm 0.3}$	86.3$_{\pm 0.2}$	86.0$_{\pm 0.3}$	87.0$_{\pm 0.2}$	84.7$_{\pm 0.4}$	86.1$_{\pm 0.2}$	83.5$_{\pm 0.7}$	85.5$_{\pm 0.3}$	85.9$_{\pm 0.2}$	83.4$_{\pm 0.3}$	81.3$_{\pm 1.3}$	65.0$_{\pm 5.2}$
Pendigits	98.7$_{\pm 1.5}$	99.5$_{\pm 0.7}$	96.7$_{\pm 4.3}$	95.0$_{\pm 4.1}$	99.4$_{\pm 0.3}$	99.5$_{\pm 0.2}$	95.9$_{\pm 5.5}$	98.2$_{\pm 1.0}$	95.8$_{\pm 3.9}$	96.2$_{\pm 3.2}$	91.5$_{\pm 13.8}$	93.1$_{\pm 3.7}$	83.1$_{\pm 8.0}$
Letter	**96.3**$_{\pm 0.9}$	95.8$_{\pm 1.3}$	90.8$_{\pm 5.6}$	89.5$_{\pm 4.3}$	94.6$_{\pm 0.9}$	95.7$_{\pm 0.9}$	90.3$_{\pm 4.3}$	63.9$_{\pm 5.9}$	67.4$_{\pm 5.1}$	62.9$_{\pm 6.7}$	72.4$_{\pm 10.7}$	54.1$_{\pm 1.4}$	62.7$_{\pm 10.4}$
Spambase	84.7$_{\pm 14.2}$	85.1$_{\pm 3.7}$	82.4$_{\pm 16.7}$	87.4$_{\pm 5.4}$	88.5$_{\pm 6.0}$	86.5$_{\pm 3.8}$	89.0$_{\pm 1.9}$	91.3$_{\pm 2.5}$	84.4$_{\pm 9.6}$	83.0$_{\pm 10.8}$	78.7$_{\pm 20.8}$	77.6$_{\pm 9.5}$	82.7$_{\pm 3.1}$
Mushrooms	99.4$_{\pm 1.7}$	98.3$_{\pm 2.5}$	99.9$_{\pm 0.2}$	99.2$_{\pm 1.6}$	99.5$_{\pm 0.6}$	99.8$_{\pm 0.1}$	99.9$_{\pm 0.1}$	99.5$_{\pm 0.3}$	99.5$_{\pm 0.8}$	97.8$_{\pm 3.9}$	93.0$_{\pm 11.0}$	95.6$_{\pm 2.4}$	95.5$_{\pm 3.4}$
Connect4	85.0$_{\pm 1.3}$	86.4$_{\pm 1.8}$	76.9$_{\pm 7.9}$	**87.9**$_{\pm 1.0}$	86.2$_{\pm 1.0}$	76.1$_{\pm 2.5}$	82.7$_{\pm 7.7}$	58.7$_{\pm 2.4}$	63.8$_{\pm 2.9}$	62.7$_{\pm 1.0}$	66.3$_{\pm 3.7}$	55.7$_{\pm 7.6}$	56.3$_{\pm 3.0}$
IMDB	81.8$_{\pm 3.6}$	82.3$_{\pm 4.2}$	62.7$_{\pm 7.0}$	80.6$_{\pm 7.3}$	79.4$_{\pm 6.5}$	77.6$_{\pm 2.0}$	64.6$_{\pm 6.9}$	84.1$_{\pm 0.4}$	**84.6**$_{\pm 0.6}$	78.5$_{\pm 2.6}$	76.3$_{\pm 4.0}$	62.8$_{\pm 5.3}$	67.3$_{\pm 4.4}$
20News	86.8$_{\pm 10.5}$	79.5$_{\pm 8.3}$	65.9$_{\pm 11.1}$	89.6$_{\pm 2.6}$	79.0$_{\pm 11.5}$	88.5$_{\pm 1.1}$	88.8$_{\pm 1.6}$	90.9$_{\pm 0.9}$	92.3$_{\pm 0.8}$	88.3$_{\pm 7.9}$	85.1$_{\pm 12.1}$	74.6$_{\pm 9.5}$	76.2$_{\pm 5.1}$
DBpedia	**99.4**$_{\pm 0.3}$	96.7$_{\pm 3.4}$	95.1$_{\pm 6.7}$	98.4$_{\pm 1.1}$	84.5$_{\pm 15.4}$	98.3$_{\pm 0.4}$	61.2$_{\pm 1.5}$	78.5$_{\pm 3.7}$	79.1$_{\pm 1.9}$	81.4$_{\pm 3.2}$	81.2$_{\pm 6.9}$	82.5$_{\pm 11.8}$	92.2$_{\pm 6.5}$
Elliptic	95.8$_{\pm 0.6}$	94.5$_{\pm 1.2}$	95.3$_{\pm 0.5}$	93.4$_{\pm 1.4}$	92.6$_{\pm 1.7}$	93.4$_{\pm 1.0}$	96.2$_{\pm 0.3}$	74.6$_{\pm 6.9}$	75.9$_{\pm 4.4}$	83.1$_{\pm 3.9}$	89.8$_{\pm 3.7}$	82.4$_{\pm 2.5}$	75.7$_{\pm 5.2}$
DiFraud	81.3$_{\pm 1.2}$	81.7$_{\pm 1.8}$	63.4$_{\pm 4.5}$	81.2$_{\pm 1.6}$	78.6$_{\pm 2.8}$	80.6$_{\pm 0.6}$	57.7$_{\pm 3.5}$	80.8$_{\pm 1.0}$	80.3$_{\pm 1.6}$	71.5$_{\pm 2.2}$	73.2$_{\pm 0.7}$	69.6$_{\pm 1.9}$	68.1$_{\pm 1.8}$
SpeechCommands	89.8$_{\pm 12.7}$	73.0$_{\pm 22.9}$	93.4$_{\pm 1.4}$	92.4$_{\pm 0.6}$	89.5$_{\pm 3.2}$	50.9$_{\pm 0.7}$	77.4$_{\pm 13.4}$	50.9$_{\pm 1.8}$	50.7$_{\pm 0.7}$	50.1$_{\pm 1.5}$	92.3$_{\pm 0.8}$	57.7$_{\pm 3.1}$	64.4$_{\pm 2.7}$
YelpChi	78.1$_{\pm 1.3}$	77.3$_{\pm 1.8}$	79.7$_{\pm 1.7}$	73.1$_{\pm 4.0}$	76.5$_{\pm 3.6}$	75.1$_{\pm 1.6}$	76.9$_{\pm 3.0}$	62.4$_{\pm 6.5}$	61.9$_{\pm 2.7}$	60.6$_{\pm 5.8}$	66.4$_{\pm 6.1}$	60.2$_{\pm 5.3}$	55.5$_{\pm 4.1}$
CICIDS2017	98.7$_{\pm 0.1}$	99.1$_{\pm 0.1}$	98.3$_{\pm 0.1}$	98.0$_{\pm 0.0}$	99.4$_{\pm 0.0}$	98.1$_{\pm 0.2}$	98.5$_{\pm 0.0}$	91.1$_{\pm 9.5}$	95.3$_{\pm 0.3}$	97.8$_{\pm 0.1}$	96.9$_{\pm 0.1}$	96.1$_{\pm 0.7}$	86.1$_{\pm 3.8}$
Rank	12.0	12.1	12.6	12.8	13.4	14.4	15.0	17.0	17.3	18.5	20.1	23.1	23.2

Supplementary F1. Our primary metric is AUC. To demonstrate practical applicability to binary decision tasks, we report supplementary test F1 under method-specific decision rules. For PURIST, DT, PseudoLabel, ELR, MeanTeacher, and SAT, we select a threshold for each dataset and seed by maximizing Youden's $J = \mathrm{TPR} - \mathrm{FPR}$ (Youden, 1950) using ground-truth labels on the held-out validation split, then apply it unchanged to the test split. This post-hoc calibration follows checkpoint selection by validation proxy AUC. In practice, a small, separately labeled calibration set containing both classes can serve this purpose. The remaining methods retain their implementation-specific decision rules.

Table 11: Supplementary test F1 (×100) (Part 1) (SCAR, c=0.1, mean±std). Decision thresholds follow the method-specific protocols in Section D.

Dataset	PN†	**PURIST**	DT	PseudoLabel	ELR	MeanTeacher	SAT	nnPUSB	PULDA	PUET	Dist-PU	nnPU
MNIST	98.2$_{\pm 0.9}$	97.0$_{\pm 0.3}$	95.7$_{\pm 1.5}$	97.5$_{\pm 0.3}$	96.5$_{\pm 1.4}$	95.0$_{\pm 2.2}$	96.6$_{\pm 0.8}$	88.4$_{\pm 3.8}$	38.3$_{\pm 46.9}$	95.6$_{\pm 0.3}$	97.7$_{\pm 0.2}$	92.5$_{\pm 7.5}$
FashionMNIST	98.7$_{\pm 0.6}$	98.0$_{\pm 0.2}$	96.9$_{\pm 1.3}$	98.0$_{\pm 0.2}$	97.8$_{\pm 0.6}$	96.5$_{\pm 2.3}$	97.8$_{\pm 0.2}$	94.7$_{\pm 1.6}$	28.7$_{\pm 43.9}$	98.0$_{\pm 0.1}$	96.8$_{\pm 1.1}$	94.9$_{\pm 3.9}$
CIFAR10	91.5$_{\pm 1.8}$	87.1$_{\pm 0.7}$	85.2$_{\pm 4.1}$	87.0$_{\pm 2.3}$	87.0$_{\pm 1.4}$	86.4$_{\pm 1.3}$	86.0$_{\pm 1.5}$	87.2$_{\pm 0.7}$	85.8$_{\pm 4.0}$	71.3$_{\pm 0.6}$	89.0$_{\pm 1.6}$	86.6$_{\pm 1.5}$
CIFAR100	44.3$_{\pm 11.0}$[5]	23.3$_{\pm 3.9}$[5]	23.0$_{\pm 1.9}$[5]	22.3$_{\pm 3.2}$[5]	29.7$_{\pm 3.4}$[5]	20.7$_{\pm 8.2}$[5]	25.1$_{\pm 5.8}$[5]	30.6$_{\pm 3.8}$[5]	39.4$_{\pm 6.4}$[5]	0.2$_{\pm 0.3}$[5]	18.6$_{\pm 14.3}$[5]	0.0$_{\pm 0.0}$[5]
SVHN	95.1$_{\pm 2.1}$	85.6$_{\pm 5.8}$	85.9$_{\pm 3.2}$	65.4$_{\pm 13.5}$	88.8$_{\pm 1.7}$	83.3$_{\pm 4.7}$	88.1$_{\pm 1.2}$	88.3$_{\pm 2.2}$	69.9$_{\pm 0.0}$	67.5$_{\pm 1.0}$	69.9$_{\pm 0.0}$	27.0$_{\pm 33.5}$
Adult	67.8$_{\pm 1.2}$	63.7$_{\pm 1.2}$	63.3$_{\pm 2.2}$	65.0$_{\pm 0.8}$	66.5$_{\pm 1.4}$	63.5$_{\pm 2.6}$	66.8$_{\pm 0.9}$	66.1$_{\pm 1.8}$	64.8$_{\pm 1.3}$	55.2$_{\pm 0.8}$	66.5$_{\pm 1.1}$	31.1$_{\pm 30.8}$
Covertype	91.6$_{\pm 0.9}$	84.9$_{\pm 0.8}$	86.7$_{\pm 1.4}$	83.0$_{\pm 0.6}$	84.9$_{\pm 1.4}$	84.5$_{\pm 0.4}$	77.6$_{\pm 0.5}$	84.3$_{\pm 0.4}$	78.9$_{\pm 0.9}$	87.8$_{\pm 0.2}$	80.8$_{\pm 0.3}$	76.0$_{\pm 1.0}$
SUSY	77.2$_{\pm 0.3}$	76.6$_{\pm 0.6}$	76.6$_{\pm 0.6}$	76.8$_{\pm 0.4}$	76.4$_{\pm 0.6}$	76.9$_{\pm 0.3}$	76.3$_{\pm 0.4}$	77.3$_{\pm 0.2}$	72.7$_{\pm 3.3}$	73.6$_{\pm 0.2}$	76.8$_{\pm 0.4}$	75.5$_{\pm 1.0}$
Pendigits	98.9$_{\pm 1.0}$	97.2$_{\pm 0.7}$	96.8$_{\pm 0.8}$	96.8$_{\pm 1.6}$	96.2$_{\pm 1.6}$	96.6$_{\pm 1.1}$	92.2$_{\pm 6.1}$	91.6$_{\pm 6.2}$	92.7$_{\pm 3.1}$	96.0$_{\pm 0.8}$	97.1$_{\pm 0.5}$	88.0$_{\pm 13.2}$
Letter	89.3$_{\pm 8.3}$	77.5$_{\pm 5.7}$	74.6$_{\pm 5.5}$	74.6$_{\pm 5.2}$	72.5$_{\pm 5.7}$	77.9$_{\pm 3.1}$	65.6$_{\pm 8.0}$	70.4$_{\pm 14.0}$	83.3$_{\pm 2.4}$	67.5$_{\pm 1.9}$	72.4$_{\pm 3.5}$	0.0$_{\pm 0.0}$
Spambase	89.4$_{\pm 2.2}$	81.7$_{\pm 3.1}$	74.5$_{\pm 6.3}$	82.2$_{\pm 4.1}$	81.0$_{\pm 7.0}$	77.2$_{\pm 7.2}$	64.3$_{\pm 25.1}$	70.8$_{\pm 12.8}$	44.6$_{\pm 36.9}$	82.2$_{\pm 2.5}$	84.0$_{\pm 2.6}$	34.4$_{\pm 38.2}$
Mushrooms	99.7$_{\pm 0.4}$	98.1$_{\pm 1.1}$	97.2$_{\pm 1.7}$	98.5$_{\pm 0.7}$	97.9$_{\pm 2.2}$	96.8$_{\pm 1.9}$	96.6$_{\pm 2.9}$	99.0$_{\pm 0.4}$	94.5$_{\pm 9.8}$	98.7$_{\pm 0.8}$	99.2$_{\pm 0.6}$	89.0$_{\pm 18.5}$
Connect4	91.2$_{\pm 1.2}$	82.3$_{\pm 1.6}$	82.3$_{\pm 1.8}$	83.1$_{\pm 1.4}$	83.5$_{\pm 2.1}$	80.1$_{\pm 3.2}$	72.8$_{\pm 7.0}$	84.6$_{\pm 1.0}$	40.8$_{\pm 40.8}$	82.5$_{\pm 0.5}$	79.4$_{\pm 0.0}$	79.4$_{\pm 0.0}$
IMDB	79.7$_{\pm 1.3}$	74.4$_{\pm 2.6}$	70.4$_{\pm 4.6}$	74.5$_{\pm 1.8}$	71.6$_{\pm 4.8}$	70.5$_{\pm 5.8}$	60.0$_{\pm 7.0}$	76.0$_{\pm 1.2}$	37.0$_{\pm 34.2}$	58.9$_{\pm 0.7}$	6.7$_{\pm 20.0}$	74.3$_{\pm 2.4}$
20News	91.6$_{\pm 0.9}$	85.3$_{\pm 2.0}$	83.4$_{\pm 1.7}$	84.8$_{\pm 1.5}$	84.4$_{\pm 2.7}$	74.4$_{\pm 10.2}$	65.0$_{\pm 10.9}$	87.4$_{\pm 0.8}$	45.4$_{\pm 38.0}$	82.1$_{\pm 0.6}$	87.3$_{\pm 1.4}$	87.5$_{\pm 0.6}$
DBpedia	92.4$_{\pm 1.3}$	71.6$_{\pm 6.3}$	62.8$_{\pm 7.7}$	70.7$_{\pm 5.9}$	72.6$_{\pm 6.8}$	56.2$_{\pm 12.5}$	65.2$_{\pm 8.0}$	83.1$_{\pm 2.5}$	78.6$_{\pm 26.7}$	0.3$_{\pm 0.3}$	87.8$_{\pm 1.2}$	0.0$_{\pm 0.0}$
Elliptic	87.3$_{\pm 2.7}$	68.5$_{\pm 6.0}$	62.1$_{\pm 6.6}$	67.5$_{\pm 8.3}$	68.5$_{\pm 5.4}$	65.7$_{\pm 11.0}$	67.4$_{\pm 4.1}$	76.6$_{\pm 3.8}$	82.5$_{\pm 1.8}$	84.2$_{\pm 1.1}$	80.8$_{\pm 1.6}$	0.0$_{\pm 0.0}$
DiFraud	75.3$_{\pm 1.2}$	70.4$_{\pm 0.9}$	67.7$_{\pm 2.2}$	69.5$_{\pm 0.9}$	67.0$_{\pm 1.6}$	66.1$_{\pm 2.7}$	52.5$_{\pm 5.6}$	67.6$_{\pm 0.6}$	71.1$_{\pm 1.1}$	36.6$_{\pm 0.7}$	33.7$_{\pm 33.7}$	56.0$_{\pm 18.8}$
SpeechCommands	79.5$_{\pm 5.0}$	70.8$_{\pm 6.7}$	61.2$_{\pm 6.5}$	39.4$_{\pm 17.0}$	74.4$_{\pm 2.7}$	66.1$_{\pm 4.4}$	69.9$_{\pm 3.5}$	71.0$_{\pm 2.2}$	68.1$_{\pm 15.5}$	10.6$_{\pm 1.5}$	17.2$_{\pm 6.7}$	0.0$_{\pm 0.0}$
YelpChi	52.0$_{\pm 10.8}$	43.2$_{\pm 2.2}$	39.9$_{\pm 4.6}$	42.9$_{\pm 2.5}$	43.0$_{\pm 3.0}$	42.1$_{\pm 3.2}$	43.9$_{\pm 1.8}$	42.3$_{\pm 6.7}$	43.1$_{\pm 2.0}$	18.6$_{\pm 1.8}$	40.0$_{\pm 1.4}$	0.0$_{\pm 0.0}$
CICIDS2017	93.3$_{\pm 0.5}$	91.4$_{\pm 0.6}$	91.5$_{\pm 0.6}$	91.1$_{\pm 0.8}$	91.3$_{\pm 0.8}$	91.3$_{\pm 0.7}$	85.3$_{\pm 1.0}$	91.8$_{\pm 0.4}$	84.8$_{\pm 1.0}$	97.8$_{\pm 0.3}$	91.9$_{\pm 0.4}$	86.2$_{\pm 1.3}$

Table 12: Supplementary test F1 (×100) (Part 2) (SCAR, c=0.1, mean±std). Decision thresholds follow the method-specific protocols in Section D.

Dataset	Self-PU	PULCPBF	VAE-PU	Holistic	LaGAM	VPU	P3Mix-E	P3Mix-C	LBE	BBE-PU	PAN	CGenPU
MNIST	95.4$_{\pm 2.2}$	91.7$_{\pm 8.7}$	68.1$_{\pm 9.1}$	94.4$_{\pm 1.9}$	86.5$_{\pm 9.8}$	94.9$_{\pm 1.9}$	87.6$_{\pm 9.7}$	93.9$_{\pm 3.0}$	11.2$_{\pm 12.3}$	95.8$_{\pm 1.3}$	94.2$_{\pm 1.1}$	72.8$_{\pm 4.8}$
FashionMNIST	93.8$_{\pm 5.0}$	91.5$_{\pm 9.2}$	69.7$_{\pm 9.5}$	97.4$_{\pm 1.0}$	96.2$_{\pm 1.1}$	95.6$_{\pm 2.7}$	93.5$_{\pm 5.6}$	94.2$_{\pm 5.6}$	16.2$_{\pm 17.2}$	96.6$_{\pm 2.4}$	92.8$_{\pm 3.6}$	73.9$_{\pm 4.2}$
CIFAR10	74.2$_{\pm 25.2}$	67.4$_{\pm 12.7}$	37.7$_{\pm 30.5}$	87.1$_{\pm 2.1}$	91.2$_{\pm 3.5}$	48.9$_{\pm 33.7}$	84.9$_{\pm 1.3}$	84.8$_{\pm 1.5}$	11.1$_{\pm 20.3}$	56.0$_{\pm 28.5}$	28.6$_{\pm 18.3}$	57.2$_{\pm 0.0}$
CIFAR100	0.0$_{\pm 0.0}$[5]	21.8$_{\pm 15.1}$[5]	0.0$_{\pm 0.0}$[5]	40.8$_{\pm 9.2}$[5]	24.7$_{\pm 3.9}$[5]	6.0$_{\pm 8.1}$[5]	17.5$_{\pm 3.1}$[5]	13.3$_{\pm 3.0}$[5]	0.0$_{\pm 0.0}$[5]	0.0$_{\pm 0.0}$[5]	5.7$_{\pm 4.7}$[5]	9.5$_{\pm 0.0}$[5]
SVHN	69.4$_{\pm 1.4}$	69.9$_{\pm 0.0}$	68.8$_{\pm 2.5}$	91.9$_{\pm 0.7}$	81.4$_{\pm 3.0}$	70.7$_{\pm 8.4}$	23.1$_{\pm 35.4}$	76.5$_{\pm 7.5}$	9.2$_{\pm 22.2}$	69.9$_{\pm 0.0}$	16.4$_{\pm 4.1}$	69.9$_{\pm 0.0}$
Adult	56.0$_{\pm 19.6}$	38.6$_{\pm 0.0}$	0.3$_{\pm 0.7}$	65.3$_{\pm 3.5}$	64.7$_{\pm 0.9}$	67.3$_{\pm 1.4}$	64.2$_{\pm 5.0}$	63.6$_{\pm 3.3}$	67.1$_{\pm 1.2}$	60.1$_{\pm 3.1}$	53.0$_{\pm 11.8}$	42.3$_{\pm 3.1}$
Covertype	76.9$_{\pm 0.7}$	65.6$_{\pm 0.0}$	53.1$_{\pm 7.5}$	80.5$_{\pm 0.4}$	80.9$_{\pm 0.8}$	73.8$_{\pm 3.2}$	82.5$_{\pm 1.5}$	80.6$_{\pm 0.6}$	1.4$_{\pm 3.3}$	76.4$_{\pm 1.2}$	74.7$_{\pm 1.8}$	65.7$_{\pm 0.2}$
SUSY	75.9$_{\pm 0.6}$	62.8$_{\pm 0.0}$	72.0$_{\pm 1.6}$	76.7$_{\pm 0.2}$	76.3$_{\pm 0.4}$	75.6$_{\pm 0.4}$	76.3$_{\pm 0.5}$	75.5$_{\pm 0.7}$	36.5$_{\pm 6.0}$	75.6$_{\pm 0.9}$	69.7$_{\pm 0.7}$	63.0$_{\pm 0.4}$
Pendigits	91.4$_{\pm 7.2}$	98.1$_{\pm 0.6}$	86.2$_{\pm 5.7}$	84.7$_{\pm 13.0}$	92.2$_{\pm 3.7}$	78.2$_{\pm 27.2}$	93.7$_{\pm 2.9}$	93.6$_{\pm 6.1}$	22.9$_{\pm 22.3}$	94.3$_{\pm 1.7}$	78.3$_{\pm 27.1}$	70.1$_{\pm 3.3}$
Letter	0.0$_{\pm 0.0}$	64.2$_{\pm 25.9}$	0.0$_{\pm 0.0}$	74.1$_{\pm 7.2}$	60.1$_{\pm 5.9}$	58.3$_{\pm 8.0}$	81.1$_{\pm 3.9}$	80.0$_{\pm 5.7}$	61.7$_{\pm 5.9}$	9.8$_{\pm 15.0}$	37.2$_{\pm 15.2}$	34.1$_{\pm 2.3}$
Spambase	33.0$_{\pm 40.8}$	54.7$_{\pm 6.0}$	14.0$_{\pm 25.1}$	57.4$_{\pm 16.9}$	79.8$_{\pm 4.7}$	44.8$_{\pm 23.2}$	42.7$_{\pm 12.5}$	45.5$_{\pm 13.3}$	59.7$_{\pm 8.2}$	76.5$_{\pm 5.2}$	52.3$_{\pm 30.1}$	71.6$_{\pm 3.5}$
Mushrooms	77.8$_{\pm 38.9}$	65.0$_{\pm 0.0}$	78.4$_{\pm 27.3}$	66.7$_{\pm 32.7}$	95.6$_{\pm 2.9}$	84.0$_{\pm 28.2}$	65.8$_{\pm 27.3}$	79.4$_{\pm 25.6}$	70.5$_{\pm 26.5}$	97.6$_{\pm 1.0}$	68.0$_{\pm 36.7}$	85.9$_{\pm 8.5}$
Connect4	79.4$_{\pm 0.0}$	79.4$_{\pm 0.0}$	76.6$_{\pm 8.3}$	83.0$_{\pm 3.6}$	76.1$_{\pm 2.4}$	84.0$_{\pm 2.4}$	75.1$_{\pm 9.4}$	85.4$_{\pm 1.3}$	0.4$_{\pm 0.9}$	79.4$_{\pm 0.0}$	79.4$_{\pm 0.0}$	79.1$_{\pm 0.4}$
IMDB	26.7$_{\pm 32.7}$	66.7$_{\pm 0.2}$	37.0$_{\pm 19.4}$	74.7$_{\pm 5.3}$	74.9$_{\pm 1.0}$	53.2$_{\pm 27.5}$	40.7$_{\pm 30.9}$	62.2$_{\pm 23.5}$	3.0$_{\pm 8.9}$	73.9$_{\pm 3.1}$	66.7$_{\pm 0.0}$	68.9$_{\pm 2.0}$
20News	67.7$_{\pm 34.1}$	72.2$_{\pm 0.0}$	64.4$_{\pm 21.5}$	78.2$_{\pm 5.8}$	82.0$_{\pm 13.7}$	53.6$_{\pm 37.2}$	21.1$_{\pm 23.0}$	33.8$_{\pm 29.0}$	55.6$_{\pm 6.5}$	82.0$_{\pm 1.8}$	62.3$_{\pm 31.5}$	75.4$_{\pm 3.2}$
DBpedia	0.0$_{\pm 0.0}$	13.5$_{\pm 0.1}$	0.1$_{\pm 0.1}$	42.6$_{\pm 36.9}$	70.2$_{\pm 5.4}$	60.8$_{\pm 31.9}$	35.9$_{\pm 35.9}$	41.5$_{\pm 28.4}$	0.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	1.3$_{\pm 4.0}$	35.6$_{\pm 16.1}$
Elliptic	0.0$_{\pm 0.0}$	75.3$_{\pm 1.4}$	0.1$_{\pm 0.3}$	65.1$_{\pm 8.8}$	52.5$_{\pm 7.4}$	57.4$_{\pm 5.5}$	71.5$_{\pm 6.3}$	69.2$_{\pm 6.8}$	73.8$_{\pm 1.4}$	4.2$_{\pm 4.7}$	40.2$_{\pm 19.2}$	20.7$_{\pm 2.7}$
DiFraud	0.0$_{\pm 0.0}$	56.0$_{\pm 0.0}$	0.0$_{\pm 0.0}$	59.1$_{\pm 12.1}$	67.2$_{\pm 0.7}$	60.7$_{\pm 20.2}$	47.7$_{\pm 22.9}$	60.2$_{\pm 10.9}$	0.0$_{\pm 0.0}$	65.3$_{\pm 4.3}$	44.8$_{\pm 22.4}$	59.9$_{\pm 1.4}$
SpeechCommands	0.0$_{\pm 0.0}$	19.2$_{\pm 15.7}$	0.7$_{\pm 1.2}$	27.5$_{\pm 14.7}$	56.8$_{\pm 9.4}$	46.7$_{\pm 8.1}$	69.1$_{\pm 12.9}$	48.4$_{\pm 24.7}$	12.1$_{\pm 20.3}$	3.2$_{\pm 9.6}$	61.9$_{\pm 6.2}$	34.7$_{\pm 1.1}$
YelpChi	0.0$_{\pm 0.0}$	35.6$_{\pm 3.8}$	2.3$_{\pm 3.6}$	30.5$_{\pm 11.8}$	40.7$_{\pm 1.9}$	32.4$_{\pm 2.9}$	42.1$_{\pm 5.2}$	39.6$_{\pm 3.5}$	37.4$_{\pm 4.1}$	1.1$_{\pm 2.7}$	17.4$_{\pm 14.5}$	25.5$_{\pm 0.2}$
CICIDS2017	84.2$_{\pm 1.4}$	74.4$_{\pm 1.5}$	29.1$_{\pm 16.9}$	88.2$_{\pm 1.3}$	89.3$_{\pm 1.8}$	86.9$_{\pm 0.5}$	89.3$_{\pm 0.8}$	86.7$_{\pm 1.4}$	85.9$_{\pm 0.2}$	44.5$_{\pm 14.2}$	83.1$_{\pm 2.1}$	47.3$_{\pm 6.3}$

D.2 PROXY-SAR-A RESULTS (INCREASING PROXY WEIGHTS)

Table 13: Proxy-SAR-A test AUC (×100) (Part 1) (c=0.1, mean±std). Best non-oracle in **bold**.

Dataset	PN†	**PURIST**	PseudoLabel	P3Mix-E	Dist-PU	WSAUC	DT	PULDA	Direct-AUC	nnPUSB	PULCPBF	ELR	LBE
MNIST	$99.8_{\pm 0.1}$	$98.3_{\pm 0.2}$	$98.2_{\pm 0.3}$	$97.8_{\pm 0.4}$	$\mathbf{99.0}_{\pm 0.4}$	$98.5_{\pm 0.9}$	$98.2_{\pm 0.4}$	$98.9_{\pm 1.0}$	$97.7_{\pm 0.7}$	$92.3_{\pm 1.3}$	$97.8_{\pm 0.7}$	$97.7_{\pm 0.5}$	$96.9_{\pm 0.8}$
FashionMNIST	$99.8_{\pm 0.1}$	$\mathbf{99.6}_{\pm 0.0}$	$99.6_{\pm 0.0}$	$99.4_{\pm 0.2}$	$99.1_{\pm 0.9}$	$99.1_{\pm 0.7}$	$99.3_{\pm 0.4}$	$99.4_{\pm 0.4}$	$99.3_{\pm 0.4}$	$98.6_{\pm 0.4}$	$99.4_{\pm 0.2}$	$99.3_{\pm 0.3}$	$99.1_{\pm 0.2}$
CIFAR10	$98.4_{\pm 0.7}$	$93.5_{\pm 0.8}$	$93.1_{\pm 0.8}$	$91.4_{\pm 0.5}$	$94.7_{\pm 0.5}$	$93.7_{\pm 0.7}$	$92.1_{\pm 1.0}$	$96.0_{\pm 0.6}$	$91.0_{\pm 1.1}$	$94.0_{\pm 0.4}$	$94.0_{\pm 1.0}$	$90.5_{\pm 1.0}$	$86.7_{\pm 2.4}$
Adult	$91.2_{\pm 0.3}$	$88.4_{\pm 0.5}$	$87.9_{\pm 0.9}$	$88.9_{\pm 0.9}$	$88.8_{\pm 0.6}$	$87.9_{\pm 0.6}$	$88.2_{\pm 0.7}$	$88.3_{\pm 1.0}$	$88.2_{\pm 0.8}$	$89.5_{\pm 0.2}$	$89.2_{\pm 0.5}$	$88.5_{\pm 1.3}$	$\mathbf{89.9}_{\pm 0.3}$
SUSY	$87.3_{\pm 0.2}$	$\mathbf{86.6}_{\pm 0.2}$	$86.6_{\pm 0.2}$	$86.3_{\pm 0.3}$	$84.5_{\pm 0.3}$	$82.6_{\pm 0.8}$	$85.7_{\pm 0.9}$	$85.9_{\pm 0.2}$	$85.1_{\pm 1.7}$	$86.0_{\pm 0.2}$	$85.0_{\pm 0.4}$	$85.5_{\pm 1.7}$	$85.9_{\pm 0.2}$
Pendigits	$99.8_{\pm 0.3}$	$97.3_{\pm 0.4}$	$96.9_{\pm 0.4}$	$96.2_{\pm 1.1}$	$96.8_{\pm 0.5}$	$95.6_{\pm 1.8}$	$\mathbf{97.4}_{\pm 0.4}$	$95.4_{\pm 5.1}$	$97.2_{\pm 0.4}$	$96.2_{\pm 2.7}$	$97.3_{\pm 0.9}$	$96.1_{\pm 0.8}$	$95.9_{\pm 2.1}$
Spambase	$96.4_{\pm 1.1}$	$94.9_{\pm 0.6}$	$94.7_{\pm 0.9}$	$91.2_{\pm 1.7}$	$94.3_{\pm 0.7}$	$94.5_{\pm 0.8}$	$94.0_{\pm 1.6}$	$84.7_{\pm 10.2}$	$94.1_{\pm 1.2}$	$90.9_{\pm 3.7}$	$91.7_{\pm 2.0}$	$94.1_{\pm 1.4}$	$92.2_{\pm 1.2}$
Mushrooms	$100.0_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.8_{\pm 0.2}$	$99.5_{\pm 0.4}$	$99.9_{\pm 0.1}$	$99.8_{\pm 0.2}$	$99.6_{\pm 0.3}$	$99.9_{\pm 0.1}$	$99.7_{\pm 0.2}$	$99.8_{\pm 0.4}$	$99.8_{\pm 0.1}$	$99.7_{\pm 0.3}$	$99.5_{\pm 1.4}$
Connect4	$91.6_{\pm 2.1}$	$\mathbf{86.6}_{\pm 0.5}$	$86.2_{\pm 0.5}$	$86.2_{\pm 0.5}$	$85.4_{\pm 0.6}$	$84.9_{\pm 0.6}$	$85.9_{\pm 1.0}$	$84.7_{\pm 1.2}$	$86.2_{\pm 0.6}$	$84.1_{\pm 1.5}$	$83.7_{\pm 0.8}$	$85.1_{\pm 1.1}$	$85.7_{\pm 0.8}$
DBpedia	$99.7_{\pm 0.1}$	$97.9_{\pm 0.6}$	$97.5_{\pm 0.4}$	$99.0_{\pm 0.1}$	$\mathbf{99.1}_{\pm 0.2}$	$98.9_{\pm 0.2}$	$97.5_{\pm 0.3}$	$99.1_{\pm 0.3}$	$98.1_{\pm 0.4}$	$98.7_{\pm 0.3}$	$98.8_{\pm 0.3}$	$94.4_{\pm 2.3}$	$61.3_{\pm 1.2}$
Elliptic	$98.0_{\pm 0.7}$	$94.5_{\pm 0.7}$	$94.2_{\pm 0.8}$	$93.6_{\pm 0.6}$	$93.3_{\pm 0.3}$	$94.4_{\pm 0.7}$	$94.0_{\pm 0.7}$	$93.2_{\pm 1.6}$	$\mathbf{94.6}_{\pm 0.7}$	$93.5_{\pm 2.0}$	$91.5_{\pm 1.0}$	$92.5_{\pm 4.2}$	$94.4_{\pm 0.5}$
DiFraud	$89.1_{\pm 0.8}$	$80.9_{\pm 0.7}$	$80.2_{\pm 0.6}$	$80.3_{\pm 2.4}$	$80.5_{\pm 0.6}$	$80.5_{\pm 0.7}$	$79.4_{\pm 0.9}$	$80.5_{\pm 0.5}$	$79.5_{\pm 0.7}$	$80.8_{\pm 0.9}$	$80.9_{\pm 0.9}$	$79.9_{\pm 1.0}$	$\mathbf{81.8}_{\pm 0.3}$
SpeechCommands	$98.1_{\pm 0.8}$	$91.4_{\pm 3.7}$	$90.8_{\pm 4.0}$	$92.2_{\pm 1.5}$	$51.8_{\pm 1.7}$	$\mathbf{94.8}_{\pm 0.6}$	$89.5_{\pm 4.9}$	$85.5_{\pm 14.0}$	$91.8_{\pm 1.6}$	$92.3_{\pm 0.5}$	$49.4_{\pm 1.3}$	$93.2_{\pm 0.6}$	$81.8_{\pm 4.6}$
YelpChi	$83.6_{\pm 2.4}$	$\mathbf{78.3}_{\pm 1.0}$	$77.9_{\pm 1.5}$	$76.2_{\pm 1.0}$	$76.6_{\pm 1.1}$	$77.2_{\pm 1.6}$	$77.2_{\pm 1.9}$	$75.1_{\pm 1.3}$	$76.2_{\pm 2.2}$	$76.4_{\pm 3.9}$	$73.9_{\pm 1.2}$	$76.5_{\pm 1.3}$	$75.8_{\pm 0.6}$
Rank	—	4.4	7.6	8.6	8.9	9.6	9.9	9.9	10.3	10.4	11.1	11.5	11.8

Table 14: Proxy-SAR-A test AUC (×100) (Part 2) (c=0.1, mean±std). Best non-oracle in **bold**.

Dataset	Holistic	SAT	LaGAM	MeanTeacher	P3Mix-C	BBE-PU	nnPU	VPU	Self-PU	PUET	PAN	VAE-PU	CGenPU
MNIST	$98.4_{\pm 0.3}$	$98.1_{\pm 0.2}$	$97.4_{\pm 0.6}$	$97.3_{\pm 1.5}$	$98.1_{\pm 0.7}$	$97.3_{\pm 0.3}$	$93.0_{\pm 1.3}$	$97.3_{\pm 1.8}$	$96.3_{\pm 0.5}$	$95.5_{\pm 0.2}$	$96.7_{\pm 0.3}$	$72.2_{\pm 10.6}$	$86.3_{\pm 5.0}$
FashionMNIST	$99.5_{\pm 0.3}$	$99.5_{\pm 0.1}$	$99.2_{\pm 0.4}$	$98.2_{\pm 1.9}$	$98.8_{\pm 1.1}$	$99.4_{\pm 0.1}$	$99.3_{\pm 0.3}$	$99.4_{\pm 0.2}$	$99.3_{\pm 0.1}$	$99.4_{\pm 0.1}$	$94.0_{\pm 6.5}$	$77.7_{\pm 8.6}$	$85.1_{\pm 3.9}$
CIFAR10	$94.8_{\pm 0.5}$	$91.4_{\pm 1.0}$	$\mathbf{96.8}_{\pm 1.9}$	$89.1_{\pm 2.6}$	$91.5_{\pm 1.1}$	$86.1_{\pm 3.0}$	$93.9_{\pm 0.7}$	$91.2_{\pm 0.7}$	$93.2_{\pm 1.1}$	$84.8_{\pm 0.4}$	$71.7_{\pm 7.2}$	$72.2_{\pm 5.1}$	$67.9_{\pm 8.4}$
Adult	$88.2_{\pm 1.7}$	$89.4_{\pm 1.0}$	$88.4_{\pm 0.8}$	$89.1_{\pm 0.9}$	$88.1_{\pm 3.2}$	$89.6_{\pm 0.4}$	$89.2_{\pm 0.5}$	$88.4_{\pm 1.2}$	$88.7_{\pm 0.6}$	$85.7_{\pm 0.5}$	$87.9_{\pm 0.9}$	$81.1_{\pm 2.1}$	$80.5_{\pm 2.3}$
SUSY	$86.5_{\pm 0.2}$	$85.7_{\pm 0.4}$	$85.7_{\pm 0.5}$	$86.0_{\pm 0.1}$	$86.4_{\pm 0.2}$	$84.5_{\pm 0.7}$	$85.9_{\pm 0.2}$	$85.4_{\pm 0.2}$	$84.4_{\pm 0.6}$	$83.6_{\pm 0.3}$	$84.7_{\pm 0.3}$	$78.9_{\pm 1.7}$	$68.5_{\pm 4.4}$
Pendigits	$97.1_{\pm 0.6}$	$95.7_{\pm 3.3}$	$96.6_{\pm 0.6}$	$97.2_{\pm 0.4}$	$95.8_{\pm 1.7}$	$95.5_{\pm 1.4}$	$94.8_{\pm 2.8}$	$94.2_{\pm 3.4}$	$96.0_{\pm 0.9}$	$94.3_{\pm 0.6}$	$91.4_{\pm 5.8}$	$88.6_{\pm 5.1}$	$84.3_{\pm 7.0}$
Spambase	$93.9_{\pm 1.0}$	$81.1_{\pm 15.0}$	$91.1_{\pm 4.0}$	$91.7_{\pm 4.0}$	$90.1_{\pm 2.0}$	$94.3_{\pm 0.7}$	$89.6_{\pm 4.7}$	$89.6_{\pm 3.3}$	$91.9_{\pm 2.6}$	$\mathbf{95.2}_{\pm 0.5}$	$81.2_{\pm 18.1}$	$84.1_{\pm 9.6}$	$84.4_{\pm 7.5}$
Mushrooms	$98.9_{\pm 0.9}$	$97.5_{\pm 3.0}$	$99.1_{\pm 1.0}$	$99.4_{\pm 0.7}$	$95.3_{\pm 6.1}$	$99.0_{\pm 0.4}$	$99.5_{\pm 0.7}$	$98.5_{\pm 1.4}$	$99.5_{\pm 0.4}$	$\mathbf{99.9}_{\pm 0.1}$	$95.9_{\pm 6.6}$	$95.3_{\pm 2.5}$	$96.3_{\pm 3.0}$
Connect4	$85.3_{\pm 0.9}$	$85.0_{\pm 0.5}$	$83.7_{\pm 1.1}$	$86.2_{\pm 0.4}$	$85.9_{\pm 0.7}$	$85.4_{\pm 0.1}$	$83.1_{\pm 3.0}$	$84.6_{\pm 2.2}$	$83.1_{\pm 1.1}$	$81.4_{\pm 0.3}$	$84.4_{\pm 0.9}$	$59.7_{\pm 6.9}$	$68.6_{\pm 3.2}$
DBpedia	$98.4_{\pm 0.3}$	$98.0_{\pm 0.9}$	$98.7_{\pm 0.6}$	$86.0_{\pm 18.2}$	$99.0_{\pm 0.5}$	$90.9_{\pm 7.9}$	$88.6_{\pm 4.0}$	$97.7_{\pm 1.0}$	$85.3_{\pm 1.8}$	$94.6_{\pm 0.4}$	$80.4_{\pm 1.6}$	$90.7_{\pm 1.5}$	$90.7_{\pm 7.9}$
Elliptic	$79.6_{\pm 16.0}$	$93.7_{\pm 0.7}$	$92.5_{\pm 1.0}$	$90.2_{\pm 1.8}$	$93.3_{\pm 0.9}$	$89.3_{\pm 1.5}$	$89.7_{\pm 2.3}$	$91.2_{\pm 1.2}$	$91.7_{\pm 1.1}$	$92.3_{\pm 0.5}$	$89.9_{\pm 2.7}$	$81.7_{\pm 7.1}$	$74.8_{\pm 9.1}$
DiFraud	$80.2_{\pm 1.3}$	$79.1_{\pm 6.1}$	$80.1_{\pm 0.7}$	$79.0_{\pm 1.0}$	$78.7_{\pm 5.0}$	$80.7_{\pm 0.8}$	$80.9_{\pm 0.4}$	$79.3_{\pm 2.8}$	$79.5_{\pm 1.1}$	$70.6_{\pm 0.5}$	$73.5_{\pm 0.9}$	$72.8_{\pm 2.7}$	$74.6_{\pm 2.1}$
SpeechCommands	$50.0_{\pm 0.1}$	$81.6_{\pm 15.8}$	$84.8_{\pm 3.5}$	$86.4_{\pm 2.5}$	$75.4_{\pm 20.7}$	$50.4_{\pm 2.7}$	$49.4_{\pm 1.5}$	$85.3_{\pm 0.7}$	$48.8_{\pm 2.1}$	$71.9_{\pm 0.3}$	$90.0_{\pm 2.5}$	$56.2_{\pm 3.6}$	$63.8_{\pm 4.1}$
YelpChi	$66.2_{\pm 12.2}$	$77.9_{\pm 1.8}$	$73.6_{\pm 2.9}$	$74.9_{\pm 3.6}$	$75.2_{\pm 1.8}$	$74.7_{\pm 1.2}$	$74.9_{\pm 1.6}$	$70.0_{\pm 3.4}$	$74.2_{\pm 1.3}$	$73.5_{\pm 1.1}$	$70.1_{\pm 5.2}$	$67.8_{\pm 3.8}$	$60.7_{\pm 3.7}$
Rank	11.9	12.4	13.1	13.8	13.8	14.0	15.5	16.1	16.6	16.8	20.5	23.3	23.3

D.3 PROXY-SAR-B RESULTS (DECREASING PROXY WEIGHTS)

Selection-shift diagnostic. Table 17 estimates the finite-pool selection TV using the first-order inclusion probabilities of the implemented fixed-size sampler. Using PN oracle AUC to estimate AUC^*, the resulting finite-sample estimate of the SAR AUC-regret bound ranges from 0.112 to 0.998, with mean 0.806 and median 0.915 across the 14 datasets. The corresponding observed PN–DT and PN–PURIST gaps are reported in the same table. Equation (9) characterizes the AUC loss induced by selection bias for the population-optimal P–U BCE scorer, weighting posterior pairs whose $e\eta$ order is reversed or tied. The observed model gaps additionally reflect finite-sample estimation and optimization.

Table 15: Proxy-SAR-B test AUC ($\times 100$) (Part 1) (c=0.1, mean$\pm$std). Best non-oracle in **bold**.

Dataset	PN†	**PURIST**	Direct-AUC	LaGAM	DT	PULDA	PseudoLabel	ELR	WSAUC	P3Mix-E	SAT	nnPUSB	PUET
MNIST	$99.9_{\pm 0.1}$	$95.2_{\pm 1.3}$	$94.6_{\pm 1.9}$	$95.5_{\pm 1.4}$	$96.3_{\pm 1.2}$	$95.0_{\pm 2.3}$	$95.6_{\pm 1.2}$	$94.9_{\pm 2.7}$	$91.5_{\pm 7.4}$	$92.1_{\pm 4.1}$	$96.7_{\pm 2.2}$	$71.5_{\pm 2.7}$	$73.6_{\pm 1.0}$
FashionMNIST	$99.8_{\pm 0.1}$	$99.0_{\pm 0.4}$	$98.6_{\pm 0.4}$	$98.4_{\pm 0.7}$	$98.8_{\pm 0.3}$	$98.6_{\pm 0.7}$	$99.0_{\pm 0.7}$	$98.7_{\pm 0.8}$	$98.8_{\pm 1.3}$	$98.3_{\pm 0.7}$	$98.5_{\pm 0.5}$	$97.2_{\pm 1.9}$	$98.6_{\pm 0.1}$
CIFAR10	$98.8_{\pm 0.5}$	$90.2_{\pm 3.3}$	$86.2_{\pm 3.1}$	$\mathbf{97.2}_{\pm 2.6}$	$85.8_{\pm 4.2}$	$94.5_{\pm 1.5}$	$86.0_{\pm 6.9}$	$81.7_{\pm 3.2}$	$86.4_{\pm 2.9}$	$74.8_{\pm 3.1}$	$86.6_{\pm 3.7}$	$88.5_{\pm 1.7}$	$53.9_{\pm 1.3}$
Adult	$91.1_{\pm 0.2}$	$68.1_{\pm 3.2}$	$68.9_{\pm 2.7}$	$78.8_{\pm 7.9}$	$69.0_{\pm 4.0}$	$74.1_{\pm 3.1}$	$65.3_{\pm 2.7}$	$64.6_{\pm 3.0}$	$68.9_{\pm 4.1}$	$69.8_{\pm 3.3}$	$69.7_{\pm 3.3}$	$66.6_{\pm 3.4}$	$64.7_{\pm 1.3}$
SUSY	$87.5_{\pm 0.2}$	$28.2_{\pm 1.2}$	$29.0_{\pm 1.4}$	$28.2_{\pm 1.5}$	$29.0_{\pm 1.3}$	$17.0_{\pm 0.7}$	$28.3_{\pm 1.3}$	$28.2_{\pm 1.0}$	$27.7_{\pm 1.0}$	$27.1_{\pm 1.9}$	$26.4_{\pm 0.8}$	$28.2_{\pm 1.4}$	$29.3_{\pm 0.3}$
Pendigits	$99.8_{\pm 0.2}$	$97.7_{\pm 1.9}$	$98.1_{\pm 1.4}$	$90.0_{\pm 6.3}$	$97.7_{\pm 1.7}$	$98.4_{\pm 1.1}$	$96.5_{\pm 2.5}$	$95.4_{\pm 7.1}$	$88.1_{\pm 6.1}$	$97.4_{\pm 0.8}$	$95.3_{\pm 4.4}$	$87.5_{\pm 3.4}$	$77.1_{\pm 1.7}$
Spambase	$96.6_{\pm 1.1}$	$64.0_{\pm 4.6}$	$\mathbf{66.9}_{\pm 5.8}$	$66.8_{\pm 17.4}$	$64.8_{\pm 4.6}$	$60.2_{\pm 8.9}$	$62.6_{\pm 4.6}$	$65.3_{\pm 4.8}$	$64.0_{\pm 4.9}$	$59.0_{\pm 4.0}$	$56.1_{\pm 9.4}$	$57.5_{\pm 10.2}$	$66.3_{\pm 3.8}$
Mushrooms	$100.0_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$99.9_{\pm 0.2}$	$99.5_{\pm 0.3}$	$99.8_{\pm 0.3}$	$99.9_{\pm 0.1}$	$100.0_{\pm 0.1}$	$99.9_{\pm 0.1}$	$99.3_{\pm 0.8}$	$99.9_{\pm 0.2}$	$99.9_{\pm 0.1}$	$99.8_{\pm 0.3}$	$100.0_{\pm 0.0}$
Connect4	$96.4_{\pm 0.4}$	$44.2_{\pm 2.5}$	$\mathbf{54.7}_{\pm 6.9}$	$36.0_{\pm 14.6}$	$51.5_{\pm 5.0}$	$51.8_{\pm 5.2}$	$32.5_{\pm 4.4}$	$46.1_{\pm 5.2}$	$32.2_{\pm 3.9}$	$48.4_{\pm 6.3}$	$33.0_{\pm 3.9}$	$28.9_{\pm 2.7}$	$35.8_{\pm 1.0}$
DBpedia	$99.7_{\pm 0.1}$	$96.1_{\pm 2.0}$	$96.8_{\pm 2.4}$	$98.0_{\pm 2.6}$	$92.9_{\pm 1.9}$	$97.7_{\pm 1.5}$	$94.1_{\pm 2.8}$	$95.6_{\pm 4.8}$	$98.0_{\pm 0.8}$	$98.0_{\pm 1.8}$	$87.4_{\pm 6.7}$	$98.1_{\pm 0.5}$	$83.2_{\pm 0.7}$
Elliptic	$98.5_{\pm 0.3}$	$93.3_{\pm 0.6}$	$92.1_{\pm 1.0}$	$92.5_{\pm 1.5}$	$92.6_{\pm 1.0}$	$93.6_{\pm 1.9}$	$93.3_{\pm 0.8}$	$93.1_{\pm 1.1}$	$93.7_{\pm 1.0}$	$92.5_{\pm 0.9}$	$91.4_{\pm 2.1}$	$92.5_{\pm 1.6}$	$\mathbf{94.7}_{\pm 0.9}$
DiFraud	$89.6_{\pm 0.2}$	$67.8_{\pm 1.4}$	$66.1_{\pm 2.6}$	$\mathbf{77.8}_{\pm 0.7}$	$65.9_{\pm 2.3}$	$65.1_{\pm 4.2}$	$66.7_{\pm 1.5}$	$62.8_{\pm 1.5}$	$65.6_{\pm 0.9}$	$63.6_{\pm 2.3}$	$52.4_{\pm 3.3}$	$65.4_{\pm 3.4}$	$59.1_{\pm 0.5}$
SpeechCommands	$98.4_{\pm 0.4}$	$92.5_{\pm 1.8}$	$93.4_{\pm 1.2}$	$88.2_{\pm 1.5}$	$89.6_{\pm 2.4}$	$87.5_{\pm 16.8}$	$54.0_{\pm 10.7}$	$95.2_{\pm 0.4}$	$\mathbf{95.9}_{\pm 1.0}$	$86.3_{\pm 17.3}$	$94.1_{\pm 1.2}$	$92.3_{\pm 1.5}$	$71.0_{\pm 1.6}$
YelpChi	$87.0_{\pm 2.8}$	$71.9_{\pm 1.5}$	$69.4_{\pm 4.1}$	$71.5_{\pm 3.5}$	$67.3_{\pm 4.4}$	$65.5_{\pm 5.4}$	$72.6_{\pm 3.1}$	$69.8_{\pm 2.7}$	$68.0_{\pm 2.8}$	$70.9_{\pm 2.7}$	$71.1_{\pm 1.7}$	$71.1_{\pm 7.9}$	$\mathbf{75.4}_{\pm 1.6}$
Rank	—	6.6	6.7	7.9	8.1	8.4	8.9	9.4	10.5	11.0	11.7	12.4	12.5

Table 16: Proxy-SAR-B test AUC ($\times 100$) (Part 2) (c=0.1, mean$\pm$std). Best non-oracle in **bold**.

Dataset	Holistic	PULCPBF	Dist-PU	LBE	MeanTeacher	P3Mix-C	VPU	PAN	CGenPU	Self-PU	BBE-PU	nnPU	VAE-PU
MNIST	$94.5_{\pm 1.2}$	$94.7_{\pm 2.6}$	$97.2_{\pm 1.0}$	$84.3_{\pm 2.6}$	$90.6_{\pm 2.2}$	$92.9_{\pm 2.5}$	$90.3_{\pm 1.4}$	$94.0_{\pm 4.9}$	$65.1_{\pm 8.4}$	$\mathbf{97.7}_{\pm 2.0}$	$84.2_{\pm 6.2}$	$80.9_{\pm 8.8}$	$69.6_{\pm 12.5}$
FashionMNIST	$98.9_{\pm 0.3}$	$98.1_{\pm 0.9}$	$\mathbf{99.4}_{\pm 0.4}$	$96.7_{\pm 0.7}$	$98.2_{\pm 0.3}$	$98.1_{\pm 0.6}$	$98.1_{\pm 3.0}$	$91.6_{\pm 6.3}$	$73.8_{\pm 7.4}$	$88.9_{\pm 6.0}$	$92.9_{\pm 5.5}$	$87.5_{\pm 4.4}$	$78.3_{\pm 7.7}$
CIFAR10	$92.2_{\pm 1.8}$	$90.2_{\pm 3.4}$	$87.3_{\pm 1.4}$	$77.9_{\pm 7.8}$	$79.7_{\pm 9.1}$	$77.5_{\pm 2.2}$	$84.0_{\pm 2.5}$	$67.9_{\pm 8.8}$	$59.7_{\pm 7.7}$	$78.3_{\pm 7.9}$	$62.0_{\pm 14.0}$	$84.0_{\pm 2.8}$	$64.6_{\pm 14.2}$
Adult	$67.9_{\pm 3.2}$	$65.3_{\pm 2.2}$	$66.2_{\pm 1.7}$	$76.1_{\pm 3.7}$	$66.3_{\pm 3.6}$	$68.1_{\pm 2.2}$	$74.0_{\pm 3.8}$	$\mathbf{80.5}_{\pm 5.2}$	$62.4_{\pm 8.1}$	$60.0_{\pm 1.3}$	$55.2_{\pm 4.0}$	$64.1_{\pm 2.1}$	$63.0_{\pm 4.3}$
SUSY	$28.2_{\pm 0.6}$	$30.0_{\pm 1.4}$	$28.8_{\pm 1.6}$	$24.6_{\pm 1.2}$	$27.8_{\pm 0.9}$	$26.0_{\pm 1.5}$	$23.4_{\pm 0.6}$	$16.6_{\pm 0.4}$	$\mathbf{39.9}_{\pm 3.6}$	$26.6_{\pm 1.4}$	$26.6_{\pm 1.5}$	$26.2_{\pm 0.6}$	$28.1_{\pm 3.9}$
Pendigits	$95.8_{\pm 1.8}$	$95.3_{\pm 2.0}$	$85.6_{\pm 1.7}$	$\mathbf{98.7}_{\pm 0.9}$	$95.0_{\pm 4.5}$	$94.1_{\pm 4.2}$	$86.6_{\pm 6.2}$	$81.2_{\pm 16.6}$	$82.8_{\pm 9.4}$	$82.4_{\pm 5.7}$	$86.8_{\pm 7.1}$	$83.2_{\pm 4.8}$	$71.3_{\pm 7.2}$
Spambase	$64.5_{\pm 5.7}$	$64.1_{\pm 4.0}$	$57.9_{\pm 4.7}$	$64.8_{\pm 4.7}$	$64.9_{\pm 4.1}$	$59.4_{\pm 4.7}$	$60.6_{\pm 5.2}$	$57.9_{\pm 11.6}$	$66.0_{\pm 4.1}$	$52.5_{\pm 7.4}$	$55.9_{\pm 4.1}$	$53.3_{\pm 7.2}$	$49.6_{\pm 5.3}$
Mushrooms	$99.4_{\pm 0.9}$	$99.9_{\pm 0.1}$	$99.8_{\pm 0.2}$	$99.9_{\pm 0.3}$	$99.5_{\pm 1.3}$	$99.7_{\pm 0.7}$	$98.9_{\pm 2.1}$	$92.5_{\pm 17.0}$	$95.9_{\pm 2.2}$	$99.4_{\pm 0.5}$	$99.8_{\pm 0.2}$	$99.8_{\pm 0.4}$	$95.0_{\pm 2.8}$
Connect4	$22.1_{\pm 1.4}$	$30.4_{\pm 3.9}$	$23.5_{\pm 1.3}$	$21.0_{\pm 1.2}$	$54.2_{\pm 5.6}$	$48.1_{\pm 6.3}$	$46.5_{\pm 13.1}$	$28.0_{\pm 5.6}$	$45.1_{\pm 4.8}$	$32.4_{\pm 1.1}$	$34.0_{\pm 7.7}$	$31.1_{\pm 1.7}$	$42.4_{\pm 5.6}$
DBpedia	$96.7_{\pm 2.4}$	$92.0_{\pm 4.3}$	$96.0_{\pm 2.1}$	$61.2_{\pm 1.3}$	$81.2_{\pm 13.1}$	$96.6_{\pm 4.3}$	$\mathbf{98.4}_{\pm 2.0}$	$87.2_{\pm 7.8}$	$88.8_{\pm 10.4}$	$73.7_{\pm 3.5}$	$68.6_{\pm 2.3}$	$71.5_{\pm 1.5}$	$71.5_{\pm 8.4}$
Elliptic	$84.3_{\pm 12.8}$	$89.8_{\pm 2.2}$	$91.6_{\pm 0.9}$	$93.8_{\pm 0.9}$	$88.1_{\pm 3.2}$	$91.6_{\pm 1.7}$	$89.8_{\pm 2.8}$	$87.2_{\pm 2.5}$	$72.0_{\pm 7.2}$	$63.4_{\pm 3.4}$	$65.3_{\pm 6.0}$	$60.7_{\pm 2.5}$	$75.0_{\pm 4.6}$
DiFraud	$63.8_{\pm 2.9}$	$61.1_{\pm 1.6}$	$50.0_{\pm 0.0}$	$55.7_{\pm 2.2}$	$61.3_{\pm 3.5}$	$63.6_{\pm 1.7}$	$59.0_{\pm 0.7}$	$60.5_{\pm 1.4}$	$57.0_{\pm 3.5}$	$56.3_{\pm 2.6}$	$61.3_{\pm 5.1}$	$55.0_{\pm 1.8}$	$53.0_{\pm 2.5}$
SpeechCommands	$53.4_{\pm 5.9}$	$54.5_{\pm 12.9}$	$53.0_{\pm 5.8}$	$76.6_{\pm 9.6}$	$88.9_{\pm 2.7}$	$63.4_{\pm 20.6}$	$91.9_{\pm 0.4}$	$92.5_{\pm 1.0}$	$62.7_{\pm 4.7}$	$50.4_{\pm 1.4}$	$50.4_{\pm 1.8}$	$49.9_{\pm 1.7}$	$57.6_{\pm 2.6}$
YelpChi	$63.1_{\pm 5.2}$	$66.9_{\pm 1.5}$	$70.1_{\pm 2.4}$	$73.5_{\pm 2.8}$	$68.5_{\pm 3.9}$	$69.5_{\pm 4.1}$	$62.8_{\pm 6.6}$	$57.2_{\pm 5.0}$	$53.8_{\pm 2.5}$	$53.2_{\pm 3.0}$	$52.6_{\pm 2.1}$	$52.8_{\pm 4.0}$	$56.2_{\pm 3.8}$
Rank	12.6	12.9	13.4	13.4	13.5	13.9	14.2	17.6	17.9	19.7	19.9	20.8	21.1

Table 17: Finite-pool SAR-B diagnostic, averaged over 10 seeds. The estimated theoretical AUC-regret bound is $\widehat{B}_{\mathrm{SAR}} = \min\{2\widehat{\mathrm{TV}}/(1-\pi), 2\,\mathrm{AUC}_{\mathrm{PN}} - 1\}$, where $\widehat{\mathrm{TV}}$ uses Monte Carlo first-order inclusion probabilities and PN estimates AUC^*. AUC gaps are PN oracle AUC minus method AUC; all quantities use the $[0, 1]$ scale.

Dataset	$\widehat{B}_{\mathrm{SAR}}$	PN$-$DT	PN$-$PURIST
Adult	0.822	0.222	0.230
CIFAR10	0.976	0.130	0.086
Connect4	0.927	0.449	0.522
DBpedia	0.994	0.068	0.036
DiFraud	0.792	0.237	0.218
Elliptic	0.970	0.059	0.052
FashionMNIST	0.903	0.010	0.009
MNIST	0.998	0.036	0.047
Mushrooms	0.112	0.002	0.000
Pendigits	0.996	0.021	0.021
SUSY	0.751	0.586	0.593
Spambase	0.932	0.318	0.326
SpeechCommands	0.524	0.089	0.059
YelpChi	0.589	0.197	0.151
Mean	0.806	0.173	0.168

D.4 PROXY-SAR-C RESULTS (PROXY-ORTHOGONAL PROPENSITY)

Table 18: Proxy-SAR-C test AUC ($\times 100$) (Part 1) (c=0.1, mean$\pm$std). Best non-oracle in **bold**.

Dataset	PN†	**PURIST**	PseudoLabel	P3Mix-E	ELR	WSAUC	PULDA	nnPUSB	PUET	Dist-PU	Holistic	Direct-AUC	DT
MNIST	$99.9_{\pm 0.1}$	$99.5_{\pm 0.1}$	$99.6_{\pm 0.1}$	$99.2_{\pm 0.3}$	$99.4_{\pm 0.2}$	$98.8_{\pm 1.1}$	$99.5_{\pm 0.3}$	$95.3_{\pm 1.9}$	$99.2_{\pm 0.1}$	$\mathbf{99.8}_{\pm 0.0}$	$98.8_{\pm 0.4}$	$99.3_{\pm 0.2}$	$98.9_{\pm 0.7}$
FashionMNIST	$99.9_{\pm 0.1}$	$99.6_{\pm 0.1}$	$99.6_{\pm 0.1}$	$99.4_{\pm 0.3}$	$99.3_{\pm 0.3}$	$99.5_{\pm 0.3}$	$99.5_{\pm 0.2}$	$98.9_{\pm 0.2}$	$\mathbf{99.7}_{\pm 0.0}$	$99.2_{\pm 0.7}$	$99.6_{\pm 0.3}$	$99.3_{\pm 0.2}$	$99.2_{\pm 0.5}$
CIFAR10	$98.9_{\pm 0.5}$	$95.9_{\pm 0.3}$	$96.1_{\pm 0.6}$	$94.7_{\pm 0.8}$	$95.5_{\pm 1.3}$	$95.8_{\pm 0.5}$	$96.2_{\pm 0.8}$	$95.5_{\pm 1.1}$	$88.1_{\pm 0.3}$	$96.5_{\pm 0.6}$	$\mathbf{97.0}_{\pm 0.6}$	$93.9_{\pm 2.0}$	$93.4_{\pm 3.0}$
Adult	$91.3_{\pm 0.3}$	$87.3_{\pm 0.6}$	$88.2_{\pm 0.3}$	$88.1_{\pm 1.6}$	$87.1_{\pm 0.7}$	$86.2_{\pm 0.7}$	$87.1_{\pm 1.3}$	$88.2_{\pm 1.2}$	$85.7_{\pm 0.4}$	$88.9_{\pm 0.6}$	$88.5_{\pm 1.6}$	$86.7_{\pm 1.4}$	$86.0_{\pm 0.9}$
SUSY	$87.5_{\pm 0.2}$	$\mathbf{86.9}_{\pm 0.2}$	$86.7_{\pm 0.6}$	$86.7_{\pm 0.1}$	$85.8_{\pm 0.9}$	$85.0_{\pm 0.6}$	$83.7_{\pm 0.8}$	$86.8_{\pm 0.2}$	$85.4_{\pm 0.1}$	$85.7_{\pm 0.2}$	$86.9_{\pm 0.2}$	$86.7_{\pm 0.3}$	$86.8_{\pm 0.2}$
Pendigits	$99.6_{\pm 0.3}$	$99.6_{\pm 0.2}$	$99.2_{\pm 0.2}$	$\mathbf{99.7}_{\pm 0.2}$	$99.1_{\pm 1.2}$	$96.8_{\pm 3.5}$	$99.2_{\pm 1.1}$	$98.1_{\pm 2.8}$	$99.4_{\pm 0.2}$	$99.2_{\pm 0.3}$	$99.4_{\pm 0.4}$	$99.4_{\pm 0.3}$	$99.3_{\pm 0.3}$
Spambase	$96.5_{\pm 1.3}$	$93.3_{\pm 1.5}$	$93.0_{\pm 1.7}$	$89.0_{\pm 2.1}$	$93.0_{\pm 1.8}$	$92.3_{\pm 1.6}$	$78.8_{\pm 13.8}$	$79.8_{\pm 6.3}$	$\mathbf{95.6}_{\pm 0.7}$	$93.8_{\pm 1.2}$	$92.3_{\pm 2.0}$	$91.8_{\pm 2.4}$	$91.6_{\pm 2.6}$
Mushrooms	$100.0_{\pm 0.0}$	$99.9_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.7_{\pm 0.3}$	$99.8_{\pm 0.2}$	$99.8_{\pm 0.3}$	$100.0_{\pm 0.0}$	$99.8_{\pm 0.4}$	$\mathbf{100.0}_{\pm 0.0}$	$99.9_{\pm 0.1}$	$98.6_{\pm 1.1}$	$99.8_{\pm 0.3}$	$99.7_{\pm 0.3}$
Connect4	$94.2_{\pm 2.4}$	$\mathbf{87.7}_{\pm 0.7}$	$86.6_{\pm 0.7}$	$86.6_{\pm 1.4}$	$86.1_{\pm 0.9}$	$84.9_{\pm 2.8}$	$83.5_{\pm 2.1}$	$85.4_{\pm 2.5}$	$83.0_{\pm 0.8}$	$70.1_{\pm 3.5}$	$82.3_{\pm 5.0}$	$85.6_{\pm 1.5}$	$85.6_{\pm 1.4}$
DBpedia	$99.7_{\pm 0.1}$	$98.5_{\pm 0.4}$	$98.3_{\pm 0.3}$	$98.8_{\pm 0.5}$	$94.7_{\pm 5.4}$	$98.7_{\pm 0.2}$	$\mathbf{99.4}_{\pm 0.3}$	$98.9_{\pm 0.3}$	$93.3_{\pm 0.5}$	$99.2_{\pm 0.3}$	$98.5_{\pm 0.7}$	$97.8_{\pm 0.5}$	$97.9_{\pm 0.3}$
Elliptic	$97.9_{\pm 0.8}$	$96.0_{\pm 0.4}$	$96.2_{\pm 0.5}$	$95.4_{\pm 0.4}$	$95.5_{\pm 0.8}$	$96.0_{\pm 0.4}$	$95.3_{\pm 1.2}$	$96.0_{\pm 0.3}$	$\mathbf{96.6}_{\pm 0.5}$	$94.0_{\pm 0.6}$	$90.6_{\pm 7.7}$	$95.5_{\pm 0.4}$	$95.0_{\pm 0.8}$
DiFraud	$89.8_{\pm 0.1}$	$\mathbf{83.2}_{\pm 0.6}$	$82.6_{\pm 0.4}$	$81.4_{\pm 1.7}$	$80.2_{\pm 0.9}$	$81.9_{\pm 1.3}$	$82.2_{\pm 0.3}$	$82.6_{\pm 0.5}$	$77.1_{\pm 0.4}$	$66.6_{\pm 16.6}$	$82.8_{\pm 0.8}$	$79.9_{\pm 1.7}$	$80.1_{\pm 2.0}$
SpeechCommands	$98.4_{\pm 0.7}$	$94.6_{\pm 1.1}$	$75.2_{\pm 18.9}$	$95.0_{\pm 0.7}$	$95.1_{\pm 0.6}$	$\mathbf{96.0}_{\pm 0.4}$	$92.9_{\pm 10.4}$	$93.7_{\pm 0.5}$	$75.7_{\pm 0.9}$	$52.7_{\pm 5.3}$	$50.7_{\pm 1.3}$	$93.1_{\pm 1.7}$	$89.1_{\pm 3.4}$
YelpChi	$86.2_{\pm 2.8}$	$\mathbf{80.2}_{\pm 0.6}$	$80.1_{\pm 0.6}$	$77.9_{\pm 1.2}$	$79.2_{\pm 0.9}$	$79.5_{\pm 1.4}$	$75.0_{\pm 5.0}$	$78.5_{\pm 3.4}$	$80.0_{\pm 0.6}$	$78.1_{\pm 1.1}$	$75.5_{\pm 1.9}$	$77.5_{\pm 2.5}$	$78.8_{\pm 1.6}$
Rank	—	4.1	6.3	8.5	9.5	9.9	10.3	10.4	10.4	10.4	10.4	11.0	12.1

Table 19: Proxy-SAR-C test AUC ($\times 100$) (Part 2) (c=0.1, mean$\pm$std). Best non-oracle in **bold**.

Dataset	LaGAM	VPU	P3Mix-C	SAT	LBE	MeanTeacher	PULCPBF	BBE-PU	nnPU	Self-PU	PAN	CGenPU	VAE-PU
MNIST	$98.0_{\pm 0.5}$	$98.7_{\pm 1.6}$	$99.0_{\pm 1.0}$	$99.1_{\pm 0.8}$	$98.3_{\pm 1.0}$	$96.2_{\pm 3.8}$	$98.3_{\pm 0.8}$	$99.5_{\pm 0.1}$	$98.2_{\pm 1.6}$	$99.2_{\pm 1.1}$	$97.9_{\pm 2.3}$	$79.4_{\pm 4.6}$	$69.0_{\pm 14.4}$
FashionMNIST	$99.4_{\pm 0.2}$	$99.3_{\pm 0.3}$	$98.7_{\pm 1.3}$	$99.5_{\pm 0.1}$	$98.8_{\pm 0.6}$	$99.3_{\pm 0.7}$	$97.9_{\pm 1.0}$	$99.6_{\pm 0.1}$	$98.1_{\pm 4.1}$	$98.3_{\pm 3.3}$	$93.5_{\pm 6.6}$	$82.6_{\pm 3.3}$	$77.7_{\pm 6.2}$
CIFAR10	$96.5_{\pm 2.0}$	$95.1_{\pm 0.9}$	$94.9_{\pm 0.8}$	$94.2_{\pm 1.5}$	$85.9_{\pm 3.8}$	$93.2_{\pm 4.0}$	$95.3_{\pm 0.6}$	$82.1_{\pm 9.6}$	$95.7_{\pm 0.8}$	$94.6_{\pm 2.2}$	$71.6_{\pm 7.2}$	$70.6_{\pm 4.5}$	$70.0_{\pm 10.3}$
Adult	$89.3_{\pm 0.5}$	$89.8_{\pm 0.7}$	$87.2_{\pm 2.2}$	$90.0_{\pm 0.4}$	$\mathbf{90.1}_{\pm 0.4}$	$86.6_{\pm 2.0}$	$89.0_{\pm 0.6}$	$88.2_{\pm 1.0}$	$88.7_{\pm 1.1}$	$89.3_{\pm 0.4}$	$86.8_{\pm 2.7}$	$74.7_{\pm 3.9}$	$78.3_{\pm 2.9}$
SUSY	$86.0_{\pm 0.5}$	$86.1_{\pm 0.1}$	$86.4_{\pm 0.3}$	$86.1_{\pm 0.3}$	$86.1_{\pm 0.2}$	$86.9_{\pm 0.2}$	$83.8_{\pm 0.6}$	$83.1_{\pm 1.4}$	$85.4_{\pm 0.5}$	$85.8_{\pm 0.2}$	$83.4_{\pm 0.3}$	$61.9_{\pm 6.8}$	$79.6_{\pm 1.9}$
Pendigits	$98.6_{\pm 0.9}$	$97.5_{\pm 2.7}$	$99.3_{\pm 0.8}$	$98.3_{\pm 2.8}$	$99.3_{\pm 0.6}$	$99.3_{\pm 0.3}$	$99.6_{\pm 0.2}$	$98.6_{\pm 0.3}$	$96.9_{\pm 3.5}$	$97.0_{\pm 3.1}$	$97.5_{\pm 1.1}$	$86.6_{\pm 4.1}$	$90.9_{\pm 5.6}$
Spambase	$91.7_{\pm 4.0}$	$89.8_{\pm 4.0}$	$89.7_{\pm 3.7}$	$83.4_{\pm 16.0}$	$90.1_{\pm 2.6}$	$90.5_{\pm 4.4}$	$89.3_{\pm 3.0}$	$92.1_{\pm 2.2}$	$79.0_{\pm 6.0}$	$77.8_{\pm 6.5}$	$74.9_{\pm 23.8}$	$85.3_{\pm 2.4}$	$73.2_{\pm 11.2}$
Mushrooms	$99.4_{\pm 0.7}$	$99.3_{\pm 0.9}$	$96.4_{\pm 5.4}$	$98.3_{\pm 1.9}$	$99.9_{\pm 0.1}$	$98.0_{\pm 4.9}$	$99.8_{\pm 0.2}$	$99.4_{\pm 0.6}$	$99.4_{\pm 0.6}$	$96.7_{\pm 6.9}$	$92.8_{\pm 11.3}$	$98.2_{\pm 1.5}$	$94.6_{\pm 3.2}$
Connect4	$81.8_{\pm 1.5}$	$84.5_{\pm 4.4}$	$85.8_{\pm 1.7}$	$77.3_{\pm 5.3}$	$86.6_{\pm 0.6}$	$85.9_{\pm 0.8}$	$76.4_{\pm 3.0}$	$64.6_{\pm 8.4}$	$64.2_{\pm 1.6}$	$63.3_{\pm 1.0}$	$67.2_{\pm 2.9}$	$59.7_{\pm 5.2}$	$56.9_{\pm 7.0}$
DBpedia	$98.7_{\pm 0.8}$	$98.7_{\pm 0.4}$	$96.6_{\pm 3.9}$	$96.8_{\pm 1.8}$	$61.7_{\pm 1.1}$	$91.4_{\pm 14.6}$	$98.4_{\pm 0.3}$	$74.2_{\pm 4.6}$	$79.7_{\pm 1.4}$	$80.8_{\pm 1.9}$	$81.1_{\pm 5.0}$	$89.8_{\pm 8.5}$	$84.4_{\pm 9.7}$
Elliptic	$93.9_{\pm 0.9}$	$93.7_{\pm 0.5}$	$95.0_{\pm 0.8}$	$95.0_{\pm 0.5}$	$96.0_{\pm 0.3}$	$92.7_{\pm 2.1}$	$93.6_{\pm 1.1}$	$73.1_{\pm 8.7}$	$78.4_{\pm 6.2}$	$84.2_{\pm 5.9}$	$89.4_{\pm 5.1}$	$76.8_{\pm 4.7}$	$77.2_{\pm 9.5}$
DiFraud	$79.6_{\pm 0.7}$	$80.7_{\pm 1.8}$	$82.5_{\pm 0.7}$	$64.6_{\pm 3.0}$	$56.0_{\pm 4.3}$	$79.5_{\pm 3.1}$	$80.4_{\pm 0.9}$	$80.8_{\pm 1.7}$	$81.2_{\pm 0.5}$	$71.1_{\pm 1.8}$	$72.7_{\pm 1.2}$	$70.2_{\pm 1.5}$	$70.7_{\pm 2.0}$
SpeechCommands	$87.6_{\pm 4.3}$	$91.9_{\pm 2.1}$	$67.6_{\pm 21.5}$	$93.4_{\pm 1.5}$	$80.4_{\pm 13.3}$	$89.3_{\pm 2.0}$	$50.4_{\pm 0.4}$	$50.8_{\pm 1.5}$	$50.4_{\pm 0.7}$	$50.8_{\pm 0.4}$	$87.5_{\pm 11.7}$	$63.3_{\pm 3.3}$	$56.6_{\pm 1.4}$
YelpChi	$76.4_{\pm 3.1}$	$73.8_{\pm 1.7}$	$78.0_{\pm 1.5}$	$79.3_{\pm 1.6}$	$76.5_{\pm 3.3}$	$77.8_{\pm 3.6}$	$74.8_{\pm 1.9}$	$63.2_{\pm 4.7}$	$61.2_{\pm 5.4}$	$61.6_{\pm 4.5}$	$70.6_{\pm 4.0}$	$56.5_{\pm 3.6}$	$63.9_{\pm 3.2}$
Rank	12.4	12.6	12.7	12.8	13.1	14.3	14.4	16.0	18.3	18.5	20.6	22.6	23.2

E FULL COMPONENT ABLATION

For statistical inference, seed results are first averaged within each dataset, and the 21 dataset means are treated as paired units. The interaction is the difference between the SAEM effect under balanced and natural P:U sampling. Confidence intervals resample datasets rather than individual seeds.

Table 20: Full two-by-two component ablation (SCAR, c=0.1, mean±std of $100 \times$ AUC). The two factors are natural versus balanced P:U sampling and hard versus SAEM-refined U targets. Superscripts give the number of seeds when fewer than 10 are available.

Dataset	Unbalanced BCE	SAEM only	Balanced only	Balanced + SAEM
MNIST	98.9±0.7	99.2±0.2	99.1±0.7	99.5±0.1
FashionMNIST	99.0±1.2	98.8±1.2	99.2±0.4	99.6±0.0
CIFAR10	94.9±0.6	95.2±0.6	94.0±3.0	95.6±0.4
CIFAR100	83.1±3.3^{5}	86.8±2.9^{5}	77.5±3.7^{5}	79.0±6.2^{5}
SVHN	88.2±13.0	83.9±16.5	91.9±3.3	89.2±13.2
Adult	89.9±0.4	90.3±0.4	86.8±2.0	87.7±0.6
Covertype	86.3±1.1	86.2±0.5	92.8±1.2	91.3±0.5
SUSY	86.5±0.2	86.6±0.1	86.9±0.1	87.0±0.1
Pendigits	96.8±2.9	96.2±3.2	99.6±0.1	99.7±0.2
Letter	89.7±5.5	90.3±5.1	95.6±0.5	96.0±0.7
Spambase	84.4±12.4	84.2±12.6	89.3±3.4	93.4±1.2
Mushrooms	99.8±0.3	99.7±0.4	99.6±0.5	99.9±0.1
Connect4	84.6±1.2	84.6±1.2	85.4±1.1	87.8±0.7
IMDB	82.2±2.7	84.0±2.0	77.8±4.4	83.2±1.2
20News	88.1±3.0	90.2±2.8	89.0±1.8	91.3±1.1
DBpedia	98.9±0.2	99.0±0.1	98.0±0.3	98.7±0.2
Elliptic	95.2±0.9	95.5±1.1	95.4±0.7	96.1±0.3
DiFraud	82.0±0.8	82.8±1.1	80.1±2.1	83.4±0.7
SpeechCommands	83.8±4.3	78.6±4.5	88.0±4.4	92.6±3.2
YelpChi	80.0±2.2	80.5±2.5	76.9±3.8	80.1±0.5
CICIDS2017	98.7±0.1	98.7±0.1	99.5±0.0	99.5±0.0

Table 21: Dataset-level inference for the component ablation. Seeds are first averaged within each dataset, and the 21 dataset means are treated as paired units. We report mean AUC differences in percentage points, 95% percentile bootstrap confidence intervals over datasets, and two-sided paired Wilcoxon p-values with Holm correction across the six contrasts.

Contrast	ΔAUC (pp)	95% CI	Holm p
SAEM effect, natural P:U	0.00	[-0.83, +0.73]	0.514
Balance effect, hard targets	+0.54	[-0.81, +1.90]	0.517
SAEM effect, balanced P:U	+1.33	[+0.52, +2.17]	0.010
Balance effect, SAEM targets	+1.88	[+0.09, +3.79]	0.065
Interaction	+1.34	[+0.41, +2.48]	0.014
Joint method vs. unbalanced BCE	+1.87	[+0.59, +3.27]	0.014

F PURIST ALGORITHM AND IMPLEMENTATION DETAILS

PURIST

Input: Labeled positives $\mathcal{P}$, unlabeled $\mathcal{U}$, minimum round size B_{floor}, EMA floor $\gamma_{\min}$, early-stopping patience p

Initialize θ; set $\varphi \leftarrow \mathbf{0}$

for epoch $t = 0, 1, 2, \ldots$ **do**

1. **Target refresh** ($t > 0$)**:** $\varphi_j \leftarrow (1-\gamma_t)\varphi_j + \gamma_t \sigma(f_\theta(x_j))\ \forall j \in \mathcal{U}$

2. **Model update:** for each WOR round, combine $\mathcal{P}$ with $n_{\text{neg}} = \min\{|\mathcal{U}|, \max(n_P, B_{\text{floor}} - n_P)\}$ U samples, train on mini-batches via Equation (2)

3. **Checkpoint and stop:** with observed validation positives, save the model whenever validation proxy AUC improves by more than δ; stop after p epochs without such improvement. Otherwise, stop after 40 epochs

end for

Return: the best saved checkpoint, or the final model when validation has no observed positives

Algorithm 1: Complete PURIST training procedure.

Unlabeled-target predictions are computed in evaluation mode with gradients disabled.

Balanced rounds. Each round uses all n_P labeled positives and $n_{\text{neg}} = \min\{|\mathcal{U}|, \max(n_P, B_{\text{floor}} - n_P)\}$ samples drawn from $\mathcal{U}$ without replacement, followed by random shuffling into mini-batches. When n_P is large, $n_{\text{neg}} = n_P$ gives 1:1 balance. When labeled positives are scarce, B_{floor} maintains a usable round size and $pos_weight = n_{\text{neg}}/n_P$ compensates the changed ratio. Each epoch contains $\lceil |\mathcal{U}|/n_{\text{neg}} \rceil$ rounds, approximately covering $\mathcal{U}$.

BCE sampling gradients. Fix the parameters θ and targets z_i, and assume that each $f_\theta(x_i)$ is evaluated independently of the other examples in its batch. Write $g_i = \nabla_\theta \ell_{\text{BCE}}(f_\theta(x_i), z_i)$. For a uniformly sampled size-b subset B of N examples, $\Pr(i \in B) = b/N$, hence

$$\mathbb{E}_B\left[\frac{1}{b}\sum_{i\in B} g_i\right] = \frac{1}{b}\sum_{i=1}^{N} \Pr(i \in B) g_i = \frac{1}{N}\sum_{i=1}^{N} g_i.$$

Thus uniform sampling without replacement gives an unbiased gradient; sampling with replacement gives the same expectation. Applying the identity separately to P and U gives the corresponding balanced objective in Equation (2). Historical SAEM targets can be conditioned on and frozen before this sampling. The result concerns the frozen-target P–U objective; it does not assert stepwise conditional unbiasedness along a random-reshuffling trajectory or cover batch-dependent forward passes such as training-mode BatchNorm.

Batchwise risk truncation. At fixed parameters θ, let $\widehat{r}_B(\theta)$ be a mini-batch estimate of the full-sample empirical negative-risk term $\widehat{r}(\theta)$, with $\mathbb{E}_B \widehat{r}_B = \widehat{r}$. Assume pointwise forward passes and a sampling law independent of θ. Writing $[a]_+ = \max(0, a)$, away from zero-risk points and when differentiation commutes with expectation,

$$\mathbb{E}_B \nabla_\theta [\widehat{r}_B(\theta)]_+ = \nabla_\theta \mathbb{E}_B [\widehat{r}_B(\theta)]_+ \quad \text{need not equal} \quad \nabla_\theta [\widehat{r}(\theta)]_+.$$

For example, let $\widehat{r}_B(\theta)$ equal $\theta + 1$ or $\theta - 1$ with equal probability. At $\theta = 1/4$, the expected gradient after batchwise truncation is $1/2$, whereas the gradient of the full-sample truncation is 1. Thus the bias is relative to the full-sample corrected risk, while the stochastic gradient remains unbiased for the expected batchwise objective. This observation concerns the risk objectives; optimizer-specific correction steps require a separate analysis.

Default hyperparameters. PURIST uses Adam with learning rate 10^{-3} and weight decay 10^{-4}, $B_{\text{floor}} = 256$, $\gamma_{\min} = 0.1$, and early-stopping patience $p = 10$ on all datasets, with minimum validation improvement $\delta = 10^{-4}$. With proxy-AUC stopping, training is capped at 999 epochs; no reported PURIST run reached this limit. Mini-batch size is 256, or 64 for the audio CNN; the controlled target strategies and component ablations use the same backbone-specific batch size.

Empty validation-positive sets. When the validation split contains no observed positives, proxy AUC is undefined. We use the final model after 40 epochs for PURIST, DT, and all batch-floor variants; 50 epochs for PN; and 20 EM epochs for LBE with the PU-Bench pretraining configuration. nnPUSB and VPU use fixed budgets of 40 and 50 epochs, respectively.

Random seeds. The main seed pool is $\{2, 25, 42, 52, 99, 103, 250, 666, 777, 2026\}$. The main CIFAR100 experiments and the label-rate sweep use the first five seeds, whereas the paired target-dynamics study uses all ten seeds on every dataset, including CIFAR100. Experiment-specific seed counts and available results are reported in the corresponding tables.

G Baseline Hyperparameters

Comparison inventory. The SCAR comparison contains 26 evaluated methods, grouped by supervision and whether the evaluated implementation receives the true class prior π: the supervised PN oracle; PURIST and the controlled U-label methods DT, PseudoLabel (Lee, 2013), ELR (Liu et al., 2020), and MeanTeacher (Tarvainen & Valpola, 2017); the generic noisy-label baseline SAT (Huang et al., 2020); the prior-dependent PU methods nnPU (Kiryo et al., 2017), nnPUSB (Kato et al., 2019), Dist-PU (Zhao et al., 2022), Self-PU (Chen et al., 2020b), PUET (Wilton et al., 2022), PULDA (Jiang et al., 2023), PULCPBF (Li et al., 2024), VAE-PU (Na et al., 2020), and WSAUC (Xie et al., 2024); and the prior-free PU methods VPU (Chen et al., 2020a), HolisticPU (Wang et al., 2023), LBE (Gong et al., 2022), PAN (Hu et al., 2021), LaGAM (Long et al., 2024), CGenPU (Papič et al., 2023), the P3Mix-C/E variants (Li et al., 2022), BBE-PU (Garg et al., 2021), and Direct-AUC.

All methods use the same dataset splits, label-generation seeds, public benchmark backbones where applicable, and validation proxy-AUC checkpoint criterion when defined. Within this shared evaluation protocol, PU baselines retain their method-specific losses, schedules, batch construction, epoch budgets, and early-stopping choices from the PU-Bench implementations (Chen et al., 2026) or the source-aligned adapters described below.

Checkpoint selection. Validation proxy AUC measures separation of observed $\mathcal{P}$ from $\mathcal{U}$ on the validation split. Under SCAR, its population ranking agrees with the class posterior, although finite estimates become noisy when separation is weak. Under SAR, it instead tracks $e(x)\eta(x)$, which need not rank by $\eta(x)$. Because the two factors are generally not identifiable from PU observations (Proposition 4.1), all methods use this shared observable non-oracle rule when the metric is defined; empty validation-positive sets use the fixed budgets in Section F.

Configuration and tuning policy. Inherited baseline hyperparameters are defined by the per-method YAML files in `config/methods/` at the PU-Bench revision in Section C, including their source and recommended dataset-specific overrides. We use the source-released dataset recipe when supported and the fixed method-specific benchmark preset otherwise. Optimizers, loss coefficients, learning-rate schedules, epoch budgets, and early-stopping settings follow these presets; the common checkpoint monitor is changed from validation proxy accuracy to validation proxy AUC, including dataset-level overrides. The additional controlled baselines, SAT, Direct-AUC/WSAUC, and modality adaptations are specified below. For the audio CNN, the batch-size setting is 64, while each method retains its own P/U batch construction; the pairwise AUC baselines use 128 examples per side as specified below. SAT additionally uses the audio learning rate given below. Presets are shared across seeds and labeling regimes.

The controlled U-label baselines share the same balanced-BCE infrastructure and early-stopping patience $p = 10$ as PURIST and use the following method-specific hyperparameters:

- **Mean Teacher** (Tarvainen & Valpola, 2017): parameter EMA decay α=0.999, updated after each sampling round. At each update, the student's BatchNorm running means, variances, and batch counters are copied to the teacher, while learned affine parameters follow the EMA. The teacher recomputes U targets once per epoch in evaluation mode; final evaluation uses the student checkpoint selected by validation proxy AUC. For this adapter's CIFAR10 SAR inputs, the logistic proxy is fitted with scikit-learn 1.6.1, NumPy 1.26.4, and SciPy 1.15.2.
- **PseudoLabel** (Lee, 2013): threshold 0.5, set as the natural sigmoid decision boundary for binary classification.

- **ELR** (Liu et al., 2020): regularization strength λ=3 and temporal ensembling momentum β=0.7, following the reference implementation's example configuration.

Official-aligned SAT baseline. SAT (Huang et al., 2020) is implemented independently from PURIST using the official training kernel: SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, cosine decay, batch size 256, 200 epochs, warm-up ending at epoch 60, target EMA momentum 0.9, and confidence-normalized sample weighting. For the audio CNN, SAT uses learning rate 0.01 and batch size 64, with the same warm-up and target-update schedule. On both graph datasets, we clip the global gradient norm to 1.0 for numerical stability; all graph SAT configurations use this setting. For the PU observation model, labeled positives are observed as class 1 and unlabeled examples as class 0; as in the original SAT algorithm, targets for *all* examples are updated after warm-up.

Direct AUC baselines. Following the contaminated-set AUC formulation of WSAUC (Xie et al., 2024), Direct-AUC minimizes the logistic pairwise loss $\log(1 + \exp(f(x_U) - f(x_P)))$ over independently shuffled P–U mini-batches. The WSAUC adapter uses the same objective after removing the highest-scoring fraction of $\mathcal{U}$ at each epoch; that fraction is the true U-set positive contamination recorded by the benchmark, while the trusted P side is retained. Both use the public backbone, Adam with learning rate 10^{-3} and weight decay 10^{-4}, pair-side batch size 128, validation proxy-AUC checkpointing, and early-stopping patience $p = 10$.

P3Mix benchmark adaptation. The available P3Mix implementation (Li et al., 2022) provides image-specific mixing models for its correction (P3Mix-C) and ELR-based (P3Mix-E) variants; the other source-data types use the adapters described below. We retain pseudo-label correction for P3Mix-C, ELR consistency for P3Mix-E, the pseudo-label mixing logic, and the source/EMA update structure. CIFAR100 reuses the same-shape CIFAR mixing network; SpeechCommands uses a mixable audio CNN. Dense tabular, text, and traffic features use a method-local mixable MLP adapter. The audio adapter exposes input and intermediate-stage mixing on the CNN in Section C, with activations, pooling, and dropout grouped with their preceding convolution; its default mixes at the input.

Other modality and execution adapters. Following the PU-Bench implementation, LaGAM uses ground-truth validation labels as its meta-learning support set. On audio, HolisticPU uses the shared CNN with a two-logit head (positive class in column zero); LaGAM uses differentiable meta-parameter versions of its layers and a $65280 \to 128 \to 128$ contrastive projection head with an intermediate ReLU and output L2 normalization. HolisticPU and PUL-CPBF apply their configured vector augmentations directly to log-Mel tensors, avoiding image conversion. VAE-PU reshapes generated vectors back to the CNN input shape for classification. CGenPU retains its flattened input adapter, and PUET retains its tree ensemble. For dynamic MLPs, P3Mix's teacher and PUL-CPBF's second-stage model are materialized before their parameters are used. VAE-PU drops incomplete training batches when a P or U loader contains more than one batch, retaining the batch when the entire pool fits in it. PUET skips candidate splits with an empty child and makes a leaf if no valid candidate remains.

Graph input adapters. Direct-scoring neural methods use the sampled GraphSAGE encoder in Section C. VPU, Dist-PU, PULDA, HolisticPU, LaGAM, P3Mix-C/E, PUL-CPBF, VAE-PU, and CGenPU instead receive fixed seed-specific sampled neighborhood features, so mixing, augmentation, and generation act on features rather than node identifiers. Their GraphSAGE aggregation weights remain trainable; the input vectors contain the root and sampled neighbors' features and neighbor means, with a mask for isolated roots. These adapters retain the methods' required output heads, mixing interfaces, and meta-parameter layers. Unlike direct scoring, their neighborhoods are fixed during training. PUET is non-neural and retains its tree ensemble on the original node features, without message passing. The SAR selection proxy uses the standardized source features, so changing the classifier backbone does not change the selection rule.

VAE-PU stage-schedule adaptation. For shortened benchmark recipes that do not reach the PN-classifier stage within the training budget, we rescale the schedule to enable classifier training and validation proxy-AUC checkpoint selection. We retain the architecture, pretraining budget,

optimizer, and total main-training budget T, and scale the reference 800-epoch stage boundaries $(400, 500, 700)$ and PN pause window $(500, 600]$ by $T/800$, rounding half up. For $T = 150, 200, 400$, the boundaries become $(75, 94, 131)$, $(100, 125, 175)$, and $(200, 250, 350)$, respectively; the pause windows are $(94, 113]$, $(125, 150]$, and $(250, 300]$. Configurations whose original schedules reach classifier training retain those schedules.

H HYPERPARAMETER SENSITIVITY

Label-rate sweep. The fixed panel in Table 5 contains MNIST, FashionMNIST, CIFAR10, SVHN, Adult, Connect4, Letter, Mushrooms, Pendigits, and Spambase. The same ten datasets and five paired seeds are used at all five labeling rates.

Hyperparameter sweeps. We first evaluate $\gamma_{\min} \in \{0.01, 0.05, 0.1, 0.2\}$ on five representative datasets. This late-stage EMA-floor sweep is insensitive over the tested range: the dataset-level mean AUC varies by at most 0.05 percentage points. To isolate the sampled P:U ratio, we then set B_{floor}=0 and vary 1:k for $k \in \{0.5, 1, 2, 4, 8\}$. Average AUC across the five dataset means ranges from 0.924 to 0.926. The ordering varies by dataset: Adult benefits from larger k, whereas DiFraud peaks at k=2. We use 1:1 as the symmetric shared setting rather than tune the ratio per dataset.

For B_{floor}, we choose low-positive settings in which the floor actually changes the sampled U count and report both c and n_P. Its effect is heterogeneous: larger floors improve Adult, while DBpedia prefers smaller values.

Table 22: Sensitivity to $\gamma_{\min}$ (SCAR, c=0.1; test AUC $\times$100). * = default.

Dataset	0.01	0.05	0.1*	0.2
MNIST	99.5±0.1	99.5±0.1	99.5±0.1	99.6±0.1
CIFAR10	95.6±0.5	95.6±0.5	95.6±0.4	95.6±0.8
Adult	87.7±0.6	87.7±0.6	87.7±0.6	87.7±0.6
Elliptic	96.1±0.3	96.1±0.3	96.1±0.3	96.1±0.3
DiFraud	83.4±0.7	83.4±0.7	83.4±0.7	83.4±0.6

Table 23: Sensitivity to the sampled P:U ratio 1:k (SCAR, c=0.1; five paired seeds; test AUC $\times$100). All variants set B_{floor}=0; * marks the shared setting.

Dataset	0.5	1*	2	4	8
MNIST	99.5±0.1	99.6±0.1	99.5±0.1	98.9±0.8	98.6±0.9
CIFAR10	95.8±0.3	95.8±0.2	95.2±0.6	95.9±0.5	95.6±0.4
Adult	87.5±0.4	87.5±0.7	88.2±1.2	88.8±0.8	89.7±1.1
Elliptic	96.2±0.4	96.0±0.2	96.3±0.3	96.2±0.5	96.3±0.3
DiFraud	83.1±1.0	83.4±0.6	83.5±0.5	82.9±0.3	83.0±0.2

Table 24: Sensitivity to B_{floor} in low-positive settings (five paired seeds; mean±sample standard deviation of test AUC $\times$100). The labeling rate is chosen so that the floor is active for at least part of the sweep. * = default.

Dataset	c	n_P	0	64	128	256*	512
Adult	0.01	92	81.8±3.3	81.8±3.3	81.8±3.3	83.6±1.7	84.0±1.5
DBpedia	0.01	28	93.2±2.4	93.6±1.6	93.3±2.0	90.5±2.3	89.7±1.4
Elliptic	0.01	36	87.3±3.0	87.3±3.0	88.3±2.3	88.8±2.8	89.7±2.5

I TRAINING DYNAMICS

Figure 1 compares the mean soft targets assigned to latent hidden positives and true negatives over the first nine SAEM refreshes. Here we quantify the corresponding early target fidelity across the full SCAR panel by comparing balanced and natural-proportion P:U sampling on every dataset and seed. To relate target refinement to the supervised gradient, consider a fresh case-control U draw (X, Y) independent of the training history $\mathcal{H}$. Conditional on $\mathcal{H}$, treat $z = z(x)$ as a fixed target function, so $P(Y = 1 \mid X = x, \mathcal{H}) = \eta(x) = \eta$. The BCE logit-gradient coefficients using z and Y differ by $Y - z$. Their conditional squared distortion is $D_x(z) = \mathbb{E}[(Y - z)^2 \mid X = x, \mathcal{H}] = \eta - 2\eta z + z^2$; replacing 0 by φ therefore reduces it by $\varphi(2\eta - \varphi)$ when $0 < \varphi < 2\eta$. On the actual training U set, we directly compute the empirical relative distortion reduction using the latent labels for analysis:

$$\widehat{R}(\varphi) = 1 - \frac{\sum_{j \in \mathcal{U}} (Y_j - \varphi_j)^2}{\sum_{j \in \mathcal{U}} Y_j^2}. \tag{16}$$

Positive values indicate that the soft targets are closer to the latent labels than the all-zero U targets in this squared gradient-coefficient sense. At the first refresh, balanced sampling gives a larger $\widehat{R}$ in 203 of 210 paired seed runs and all 21 dataset means. Treating dataset means as paired units gives a two-sided Wilcoxon $p = 9.54 \times 10^{-7}$.

Table 25: Relative reduction (%) in squared U-target error at the first SAEM refresh, measured against latent labels with the initial all-zero targets as the reference. Paired runs use the same dataset and seed and differ only in P:U sampling. Values are mean±standard deviation over ten seeds for every dataset, including CIFAR100; the macro row weights datasets equally.

Dataset	Balanced	Natural	Difference
20News	54.3±1.6	7.5±0.6	+46.9
Adult	48.9±1.2	7.3±0.9	+41.6
CICIDS2017	73.3±0.6	9.8±0.7	+63.5
CIFAR10	52.1±2.0	8.3±3.0	+43.7
CIFAR100	14.7±3.4	1.3±0.8	+13.3
Connect4	55.0±1.2	8.8±2.0	+46.2
Covertype	57.5±0.4	8.5±1.0	+49.0
DBpedia	34.1±5.2	6.7±0.6	+27.4
DiFraud	43.2±2.6	6.8±1.6	+36.4
Elliptic	60.6±1.2	5.9±0.6	+54.7
FashionMNIST	66.3±2.1	9.6±2.2	+56.7
IMDB	51.4±3.2	7.6±1.5	+43.8
Letter	57.4±1.3	3.8±1.0	+53.5
MNIST	66.2±2.3	9.6±1.7	+56.7
Mushrooms	68.5±1.5	15.4±2.2	+53.1
Pendigits	66.7±2.0	8.5±2.0	+58.1
SUSY	53.4±0.6	8.5±1.3	+44.9
SVHN	47.0±1.4	6.5±2.9	+40.5
Spambase	58.5±3.3	2.1±0.7	+56.4
SpeechCommands	14.3±17.4	8.0±3.0	+6.3
YelpChi	23.4±2.0	2.6±0.2	+20.7
Macro average	50.8	7.3	+43.5

J Distributed Training and Federated Learning

This section describes the structural compatibility of PURIST's per-sample state with distributed and federated workflows.

Computational structure. After the initial hard-target epoch, training alternates two phases: (1) φ **refresh:** a forward pass over all $\mathcal{U}$ to compute $\sigma(f_\theta(x_j))$, followed by a pointwise EMA update $\varphi_j \leftarrow (1-\gamma)\varphi_j + \gamma\sigma(f_\theta(x_j))$; (2) **Model update:** standard BCE mini-batch SGD. The soft-label vector φ is a 1D float32 array requiring $4|\mathcal{U}|$ bytes. In the CICIDS2017 profile with $|\mathcal{U}|$=2,241,948, this persistent target occupies 8.97 MB; together with the latest full-U score vector used by the refresh, the two arrays occupy 17.94 MB. The refresh is evaluated in batches, and each φ_j depends only on $f_\theta(x_j)$, so both computation and state can be partitioned across samples.

Data parallelism (DP). BCE gradient synchronization follows standard data-parallel training (Narayanan et al., 2021). With fixed data sharding, each worker can retain and refresh the entries of φ associated with its local samples, so the targets themselves do not require all-gather. Implementations that reshuffle samples across workers must additionally transfer the corresponding indexed target state.

Pipeline, tensor, sequence, and expert parallelism. φ enters only when the final prediction is converted to a training target and loss. Earlier model stages therefore follow the backbone's usual parallelization scheme (Narayanan et al., 2021; Li et al., 2023; Lepikhin et al., 2021). Tensor, sequence, or expert parallelism remains possible provided that each final prediction can be matched to its persistent sample identifier and target entry.

Spark ML. In a partitioned data engine such as Spark MLlib (Meng et al., 2016), φ can be stored as keyed per-sample state and updated after the epoch-level prediction pass. The implementation must preserve stable sample identifiers across repartitioning; its exact communication cost depends on whether this state remains co-partitioned with the examples.

Federated learning (FL). Following the client–server model-aggregation workflow of federated learning (McMahan et al., 2017), each client can maintain φ_j for its own $\mathcal{U}$ samples as local state. A corresponding training loop is:

1. Server broadcasts global model θ.
2. Each client locally refreshes φ using θ, then performs local model updates on its $(\mathcal{P}, \mathcal{U}, \varphi)$.
3. Clients upload model updates; the server aggregates them and updates θ.

This construction does not directly transmit φ, although the usual privacy risks of sharing model updates remain and require the same protections as in the underlying federated protocol.

Comparison with PU baselines. PU objectives have different state and aggregation patterns: nnPU applies a nonnegative correction to empirical risk terms (Kiryo et al., 2017); LBE optimizes two networks and latent label estimates (Gong et al., 2022); HolisticPU stores prediction histories for trend tests (Wang et al., 2023); and VPU applies a logarithm to a mini-batch mean of unlabeled predictions (Chen et al., 2020a). These objectives involve method-specific state and aggregation patterns; PURIST refreshes a scalar target keyed by sample identity.

K TRAINING-TIME COMPARISON

Table 26 summarizes wall-clock training time for all 26 methods on the 21 SCAR datasets. We use ten seeds per dataset, except five for CIFAR100. Times are aggregated across experiment batches under each method's training and stopping settings, providing an approximate comparison of practical training cost.

Table 26: Relative wall-clock training time for 26 methods on 205 paired dataset–seed configurations across 21 datasets (SCAR, c=0.1). Ratios to PURIST are geometrically averaged over seeds within each dataset, then across equally weighted datasets. PURIST = 1; lower is faster. †: supervised PN reference.

Method	Time	Method	Time	Method	Time
PseudoLabel	0.88	P3Mix-E	2.04	SAT	17.05
MeanTeacher	0.88	VPU	4.19	PULDA	21.92
DT	0.88	nnPU	4.36	Holistic	23.72
PURIST	**1.00**	BBE-PU	4.77	PULCPBF	24.76
ELR	1.09	nnPUSB	5.12	CGenPU	51.37
PN†	1.30	PUET	8.35	LaGAM	66.20
Direct-AUC	1.58	PAN	9.43	Self-PU	78.85
WSAUC	1.89	LBE	11.83	VAE-PU	83.11
P3Mix-C	1.89	Dist-PU	14.72		