ADAPTIVE POPULATION-RISK RECONSTRUCTION FOR LEARNING UNDER SELECTIVE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective labels arise when outcomes are observed only for a nonrandom subset of the population, causing empirical risk minimization (ERM) to optimize prediction for the selected sample rather than the deployment population. We propose Adaptive Population-Risk Reconstruction (APR), a framework that exploits bridge observations: labeled cases that are unexpectedly selected yet resemble regions where outcomes are largely unobserved. APR combines selection surprise with affinity to the unlabeled population to construct continuous Target-relevance scores, and uses these scores to reconstruct separate Core and Target components of deployment-population risk. A global reliability term limits reconstruction when effective bridge information is weak, while an observation-specific gate restricts adaptation to inputs for which the reconstructed predictor is locally relevant. Bridge discovery is completed and frozen before outcome-model training. In progressive bridge-information simulations, APR recovers 94% of the AUC gap between ERM and a full-information oracle. Across six selective-label benchmark settings, APR achieves the highest mean Target-population AUC and the lowest mean log-loss among the compared methods. In a real-covariate semi-synthetic benchmark based on UCI Bank Marketing, APR also achieves the best mean Target- and full-population AUC, log-loss, and Brier score. Additional stress tests show robustness to bridge contamination, identification noise, moderate transportability violations, and moderate misspecification of the Target-population share, while performance degrades as bridge identification or outcome transportability approaches its limiting failure regime.

1 INTRODUCTION

Selective-label problems create a fundamental mismatch between the training and deployment populations. In many institutional settings such as loan approval, medical testing, and customer management, covariates are recorded for a broad population, but outcomes are observed only for individuals who pass through a prior decision or operational filter. Standard empirical risk minimization (ERM) therefore optimizes prediction for the selected labeled sample, not for the population on which the model will be deployed. When the deployment population contains a well-observed Core component and an under-observed Target component, this distortion is structural: ERM learns primarily from the Core distribution, and collecting more data under the same observation process reinforces the same imbalance rather than correcting it. In practice, the resulting predictor may be the least reliable precisely where better prediction is most needed, for the under-observed population segments. The central challenge is therefore not simply to obtain more labels, but to recover predictive information about a population component whose outcomes are rarely observed.

A large literature addresses the resulting distortion. Classical missing-data and sample-selection methods model or reweight the observation mechanism (Rubin, 1976; Heckman, 1979; Zadrozny, 2004; Shao & Wang, 2016). Covariate-shift and domain-adaptation approaches align source and target distributions using unlabeled covariates (Huang et al., 2006; Gu et al., 2021), while a related transportability literature studies when causal or predictive relationships transfer across populations (Bareinboim & Pearl, 2013). Other work explicitly models the label-generation mechanism and incorporates it into learning (Sportisse et al., 2023), or exploits decision-maker heterogeneity to extract information from weakly observed regions (Kleinberg et al., 2018; De-Arteaga et al., 2018). More recent approaches address selective or missing-not-at-random labels through EM-based

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.

pseudo-labeling, deep sample-selection models, bias-variance-aware correction, doubly robust estimation, and imputation correction (Chang & Wiens, 2024; Cong et al., 2026; Ha et al., 2024; Zheng et al., 2025; 2026). These methods show that both the observation mechanism and the unlabeled covariate distribution contain useful information. Yet all of these methods tend to correct for, model, or impute the missing labels. A complementary possibility remains largely unexplored, that the selected sample may already contain genuinely labeled traces of the under-observed population.

We focus on this complementary source of information, which we call *bridge observations*. Consider a loan applicant who is approved despite having characteristics that typically lead to rejection, and whose profile closely resembles the population of rejected applicants for whom outcomes are unobserved. This applicant is informative because two conditions hold simultaneously - the observed label is surprising given the selection process, and the applicant's covariates align with the under-observed population. Neither condition alone suffices: a surprising label may reflect an outlier rather than the under-observed population, and covariate similarity to unlabeled cases does not make an observed outcome informative about them.

We propose *Adaptive Population-Risk Reconstruction* (APR), a framework that uses bridge observations to reconstruct the risk of the deployment population. APR combines selection surprise with unlabeled-population affinity into a continuous Target-relevance score for each labeled observation, then uses these scores to form separate empirical approximations to the Core and Target components of population risk. The two components are combined according to the Target share in the intended deployment population. This makes APR a reconstruction of population risk, rather than only a reweighting of selected samples. APR is adaptive in two ways. A global reliability term limits the influence of the reconstructed Target component when effective bridge information is insufficient, and an observation-specific gate restricts adaptation to inputs for which the reconstructed predictor is locally relevant. Bridge discovery is completed and frozen before outcome-model training, preventing the supervised loss from redefining which observations appear Target-relevant after seeing outcomes. We also characterize when bridge-based reconstruction can and cannot succeed. We provide sufficient conditions linking bridge identification, outcome transportability, and population-mixture specification, under which the reconstructed risk converges to the deployment-population risk.

Our contributions are as follows:

1. A bridge-based formulation of selective-label learning. We formalize the notion of bridge observations and show how their joint selection-surprise and Target-alignment properties yield direct outcome evidence for the under-observed population.

2. Adaptive population-risk reconstruction. We propose APR, which converts continuous bridge relevance into separate Core and Target risk components, recombining them at deployment-population proportions with adaptive reliability and gating safeguards.

3. Sufficient conditions and identification boundary. We provide sufficient conditions for consistency of the reconstructed risk and characterize the failure mode. The inability to distinguish between the Core and Target among labeled cases renders bridge discovery impossible.

4. Empirical validation. Across progressive bridge-information simulations, APR recovers 94% of the AUC gap between ERM and a full-information oracle. On six selective-label benchmarks and a real-covariate semi-synthetic benchmark, APR achieves the best mean Target- and full-population AUC, log-loss, and Brier score among compared methods, and remains robust under stress tests for bridge contamination, identification noise, transportability violations, and Target-share misspecification.

2 Problem Formulation

Consider a population of N observations indexed by $i = 1, \ldots, N$. For each observation, X_i denotes the covariates used for prediction, Y_i is the outcome of interest, and G_i contains variables related to the label-selection process. The outcome is observed only when the selection indicator $S_i = 1$. The available data are therefore

$$\mathcal{D} = \{X_i,\, G_i,\, S_i,\, S_i Y_i\}_{i=1}^{N}, \qquad S_i \in \{0, 1\}. \tag{1}$$

Thus, X_i, G_i, and S_i are observed for every unit, while Y_i is observed only for selected units. The variables in X_i and G_i may overlap but serve different roles. X_i defines the prediction problem, whereas G_i helps model why labels are observed.

Let f be a prediction rule and $\ell(\cdot, \cdot)$ a task-specific loss. The deployment-population risk is

$$R_{\text{pop}}(f) = \mathbb{E}_{(X,Y)\sim P_{\text{pop}}} \left[\ell\{f(X), Y\}\right]. \tag{2}$$

A standard supervised learner observes outcomes only when $S = 1$, and therefore targets $\mathbb{E}[\ell\{f(X), Y\} \mid S = 1]$ which need not equal $R_{\text{pop}}(f)$ when selection is nonrandom. Additional data collected under the same selection process can reduce variance within the selected sample while leaving the deployment-population bias unchanged.

We focus on settings in which the deployment population is a mixture of a well-observed Core component and an under-observed Target component. Let $Z \in \{C, T\}$ denote the latent population component and define

$$R_g(f) = \mathbb{E}\left[\ell\{f(X), Y\} \mid Z = g\right], \qquad g \in \{C, T\}.$$

Then

$$P_{\text{pop}} = (1 - \pi_T)P_C + \pi_T P_T,\ R_{\text{pop}}(f) \quad = (1 - \pi_T)R_C(f) + \pi_T R_T(f),\ 0 \le \pi_T \le 1, \tag{3}$$

where π_T is the Target share in the intended deployment population. The component label Z is introduced only to define the population estimand. APR does not require observed component labels during training or deployment.

The selective-label setting of interest is one in which the Target component is less likely to be labeled, $\Pr(S = 1 \mid Z = T) < \Pr(S = 1 \mid Z = C)$, so the selected sample may contain substantially less information about $R_T(f)$ than about $R_C(f)$. APR addresses this imbalance by using labeled observations that are informative about the Target component despite being selected. These bridge observations are useful when they are both surprising under the selection process and covariate-aligned with under-observed regions of the population.

Bridge observations need not form a random sample from P_T. The key requirement is that Target-relevant labeled observations must be at least partially distinguishable from Core observations using observed information, and their outcome information must be sufficiently transportable to the corresponding Target population. Section 3 formalizes this idea through a continuous bridge-relevance score and uses it to reconstruct the two component risks in Eq. (3).

3 Methodology: Adaptive Population-Risk Reconstruction

APR reconstructs deployment risk from selectively labeled data by separating where labels are observed from where the model will be used. It proceeds in four stages: estimating full-sample selection propensity and unlabeled-domain affinity; combining them into a soft bridge-relevance score for each labeled observation; using these scores to reconstruct empirical Core and Target risks and train a reconstructed-risk predictor, and combining this predictor with a selected-sample predictor through a reliability-weighted adaptive gate. All auxiliary scores are estimated before outcome-model training and then held fixed, preventing the supervised loss from redefining which observations appear Target-relevant. Full estimation details, identification results, and calibration variants are provided in Appendix B.

3.1 Soft Bridge Relevance

APR begins by estimating two auxiliary models using the full sample, including both labeled and unlabeled observations. The first model estimates the probability that an outcome is observed from the selection variables G_i. The second estimates affinity with the unlabeled covariate distribution from the prediction covariates X_i:

$$\widehat{e}_i = \widehat{P}(S_i = 1 \mid G_i), \qquad \widehat{d}_i = \widehat{P}(S_i = 0 \mid X_i). \tag{4}$$

The quantity $1 - \widehat{e}_i$ measures *selection surprise* for a labeled observation. It is large when the observation received a label despite having selection characteristics associated with non-observation.

The quantity $\widehat{d}_i$ measures unlabeled-domain affinity. It is large when the observation lies in a region of covariate space where outcomes are often missing.

Neither signal is sufficient on its own. A labeled observation may be surprising under the selection model because it is an outlier, not because it carries useful Target information. Conversely, an observation may resemble unlabeled covariates while providing little evidence that its label is informative about Target outcomes. APR therefore combines the two signals into an anchored bridge score b_i. For labeled observations $S_i = 1$, define

$$b_i = \mathcal{N}\Big[(1-\widehat{e}_i)\widehat{d}_i\Big], \qquad q_i = \mathcal{S}_\tau\Big[\alpha b_i + (1-\alpha)\widetilde{d}_i\Big], \qquad S_i = 1. \tag{5}$$

Here, $\mathcal{N}(\cdot)$ is a robust monotone normalization, $\widetilde{d}_i$ is the prevalence-adjusted unlabeled affinity, and $\mathcal{S}_\tau(\cdot)$ is a monotone sharpening transformation. The parameter $\alpha \in [0,1]$ controls the relative contribution of direct bridge evidence and population affinity. Larger α places more weight on selection surprise among labeled observations, while smaller α places more weight on covariate alignment with the unlabeled population.

The resulting score $q_i \in [0,1]$ is interpreted as continuous Target relevance. It is not a posterior probability of latent Target membership. In general,

$$\Pr(S=0 \mid X) \neq \Pr(Z=T \mid X) \tag{6}$$

so $\widehat{d}_i$ should be viewed as a proxy for Target relevance, not as a direct estimate of component membership. This distinction is important as APR does not require identifying which observations are truly Target or Core. It only requires a ranking or soft weighting that enriches labeled observations carrying Target-relevant outcome information.

When reliable information about the Target-population share is available, $\widehat{d}_i$ may incorporate that share through prevalence adjustment. When it is unavailable, the Target mixture weight π_T in Equation (3) can be treated as a sensitivity parameter, and the unadjusted affinity $\widehat{d}_i$ can be used for bridge scoring. The formal results are stated in Appendix B.2, with proofs in Appendix B and corresponding diagnostic experiments in Appendix C.

All quantities in Equation (5) are computed before training the outcome predictors and are then frozen. In particular, no gradient from the outcome loss is allowed to update them. This anchoring step prevents the prediction model from assigning high relevance to observations merely because doing so reduces the current supervised loss.

3.2 Soft Population Reconstruction

APR uses the soft relevance scores q_i to construct separate empirical approximations to the Core and Target components of the deployment risk. For each labeled observation, define the unnormalized Target and Core masses

$$u_i^T = q_i^{p_T}\widehat{d}_i^{\,r_T}, \qquad u_i^C = (1-q_i)^{p_C}, \qquad S_i = 1, \tag{7}$$

where p_T, r_T, and p_C are positive concentration parameters. Larger values of p_T and r_T concentrate the Target component on observations with the strongest evidence of Target relevance and unlabeled-domain affinity. p_C controls how quickly observations lose Core mass as their Target relevance increases.

These masses produce the reconstructed empirical risk

$$\widehat{R}_{\mathrm{rec}}(f) = (1-\pi_T)\frac{\sum_{i:S_i=1} u_i^C \ell\{f(X_i), Y_i\}}{\sum_{i:S_i=1} u_i^C} + \pi_T \frac{\sum_{i:S_i=1} u_i^T \ell\{f(X_i), Y_i\}}{\sum_{i:S_i=1} u_i^T}. \tag{8}$$

The first term approximates the Core risk $R_C(f)$, while the second approximates the Target risk $R_T(f)$. The mixture weight π_T sets the intended deployment composition. If π_T is known from external population information, APR uses that value. Otherwise, Appendix J provides a method to approximate unknown π_T, and its chosen value can be varied to study sensitivity to deployment composition (See Appendix I). Equation (8) directly mirrors the decomposition in Eq. (3). Appendix B provides the formal sufficient conditions and the corresponding error decomposition.

APR trains two predictors. The baseline predictor $\widehat{f}_0$ is obtained by ordinary empirical risk minimization over the labeled sample:

$$\hat{f}_0 \in \arg\min_{f \in \mathcal{F}} \frac{1}{n_L} \sum_{i:S_i=1} \ell\{f(X_i), Y_i\}, \tag{9}$$

where $n_L = \sum_i S_i$. The reconstructed predictor $\hat{f}_R$ minimizes Equation (8)

$$\hat{f}_R \in \arg\min_{f \in \mathcal{F}} \widehat{R}_{\mathrm{rec}}(f). \tag{10}$$

Training both models is useful because reconstruction is an estimate rather than a directly observed risk. The reconstructed predictor can recover relationships that receive too little weight under selected-sample training, while the baseline predictor remains stable in regions that are well represented by labeled data.

Equation (8) also clarifies how APR differs from simple bridge reweighting. Fixed bridge reweighting only increases the influence of high-relevance observations within the selected-sample objective. APR instead reconstructs separate Core and Target risks and assigns them population-level mass. The goal is therefore not merely to emphasize rare labeled observations, but to approximate the distribution on which the predictor will be deployed.

3.3 Reliability-Gated Prediction

A reconstructed Target component should not dominate predictions merely because a few labeled observations receive high Target relevance. APR therefore estimates a global reliability score that reflects both the amount and concentration of available Target information. Using the Target masses from Eq. (7), define

$$\mathrm{ESS}_T = \frac{\left(\sum_{i:S_i=1} u_i^T\right)^2}{\sum_{i:S_i=1} (u_i^T)^2}, \qquad c_T = \frac{\sum_{i:S_i=1} u_i^T q_i}{\sum_{i:S_i=1} u_i^T}, \qquad \gamma = c_T \left[1 - \exp\left(-\frac{\mathrm{ESS}_T}{\kappa}\right)\right]. \tag{11}$$

where ESS_T is the effective sample size of the reconstructed Target component, c_T is the average relevance of the Target-weighted observations, and $\kappa > 0$ controls how quickly reliability increases with effective Target information. Thus, $\gamma \in [0, 1]$ is large only when Target information is both relevant and sufficiently abundant. If the Target mass is concentrated on only a few points, or if the Target-weighted observations have weak relevance, γ remains small.

Global reliability does not determine whether reconstruction is relevant for a particular new observation. We therefore estimate a soft population gate. Let $\widetilde{a}(X) \in [0, 1]$ denote the gate probability before sharpening. Rather than training the gate on hard bridge indicators, APR uses q_i as a soft target:

$$\mathcal{L}_{\mathrm{gate}} = -\sum_{i:S_i=1} \left[q_i \log \widetilde{a}(X_i) + (1 - q_i) \log\{1 - \widetilde{a}(X_i)\}\right], \tag{12}$$

with class balancing applied when Target-relevant observations are strongly imbalanced. The fitted gate is sharpened to obtain $\widehat{a}(x)$, with tuning parameter ζ controlling the strength of local adaptation (see Appendix B.4 for more details).

For binary outcomes, let $\widehat{p}_0(x)$ and $\widehat{p}_R(x)$ denote the probabilities produced by the base and reconstructed predictors. APR combines them as

$$\widehat{p}^{\,\mathrm{raw}}(x) = [1 - \gamma \widehat{a}(x)]\, \widehat{p}_0(x) + \gamma \widehat{a}(x) \widehat{p}_R(x). \tag{13}$$

This prediction separates two roles. The global reliability γ determines how much the reconstructed predictor can be trusted overall. The local gate $\widehat{a}(x)$ determines where reconstruction is relevant. A large local gate alone cannot force strong adaptation when Target information is unreliable, and a large global reliability cannot force adaptation for observations that appear Core-like.

The limiting cases are intuitive. When Target information is scarce, γ is close to zero and APR remains near the ordinary selected-sample predictor. When an observation is strongly Core-like, $\widehat{a}(x) \approx 0$ and the prediction remains close to $\widehat{p}_0(x)$. APR assigns substantial weight to the reconstructed predictor only when Target information is globally reliable and the new observation is locally Target-relevant.

At deployment, APR requires only the task covariates X. The selection indicator S, selection variables G, and latent component identity Z are not required for prediction. Algorithm A summarizes the complete APR procedure.

4 Recovery toward Full Information

We first test APR in a setting where its intended behavior can be examined directly: as valid bridge information becomes more abundant, does the feasible estimator approach a predictor trained with full outcome information? The simulation population contains Core and Target components with different outcome mechanisms. Core outcomes are well observed, while Target outcomes are mostly unobserved except for selected Target cases that serve as valid bridges. Across seven regimes, the number of observed Target bridge labels increases from 40 to 15,000, raising the bridge share among labeled observations from about 2% to 11%, while the underlying Core–Target deployment population is fixed. ERM therefore receives more Target labels across regimes, but the labeled sample remains Core-dominated.

We compare five estimators. Unadjusted ERM trains on the observed labeled sample with equal weights. Fixed Bridge Reweighting uses the same bridge-relevance scores as APR to upweight bridge-like observations, but does not reconstruct separate Core and Target risk components. APR uses feasible bridge scores, reliability weighting, and local gating. Oracle Population-Risk Reconstruction uses the true identities of labeled bridge observations when reconstructing population risk, but still does not observe the remaining unlabeled Target outcomes. Full-Information Oracle observes all Core and Target outcomes and provides the reference ceiling. All models that do not directly observe π_T must estimate it using the approach described in Appendix J. Performance is evaluated on Test A, containing only Target observations, and Test B–Full, representing the full Core–Target deployment population.

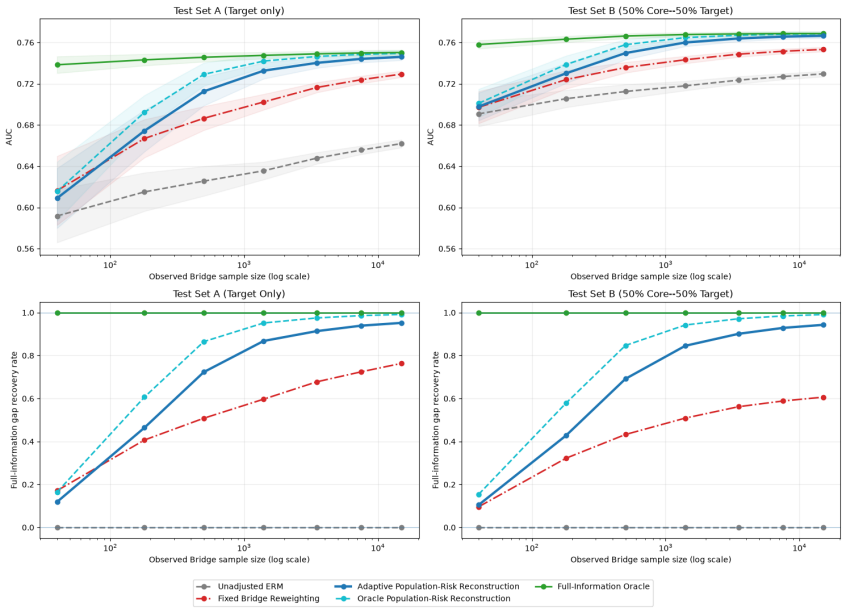


Figure 1: Recovery as observed bridge information increases. The upper panels report AUC on Test Set A (Target only) and Test Set B (50% Core–50% Target). The lower panels report the corresponding fraction of the ERM–full-information AUC gap recovered by each estimator.

To compare recovery across bridge-information regimes, we report the fraction of the ERM–full-information AUC gap recovered by method m,

$$\mathcal{R}_m = \frac{\mathrm{AUC}_m - \mathrm{AUC}_{\mathrm{ERM}}}{\mathrm{AUC}_{\mathrm{FI}} - \mathrm{AUC}_{\mathrm{ERM}}}, \tag{14}$$

where $\mathrm{AUC}_{\mathrm{FI}}$ denotes the Full-Information Oracle. $R_m = 0$ matches ERM, while $R_m = 1$ closes the full ERM-to-oracle gap. Results are averaged over 30 Monte Carlo replications, with recovery rates computed within each replication before averaging. The shaded regions in Figure 1 show ± 1 Monte Carlo standard deviation.

Figure 1 shows a clear recovery path. ERM improves gradually as more Target labels enter the observed sample, while Fixed Bridge Reweighting provides a larger but incomplete gain. Oracle Population-Risk Reconstruction moves toward the Full-Information Oracle earlier, reflecting its access to true labeled bridge identities. APR follows the same trajectory using only feasible bridge scores: it remains conservative when bridge information is scarce, but increasingly approaches the oracle benchmarks as bridge evidence becomes reliable. On Test A, APR's AUC rises from about 0.61 to 0.7460, compared with 0.7495 for Oracle Population-Risk Reconstruction and 0.7502 for the Full-Information Oracle. This corresponds to 95.2% recovery of the ERM-to-full-information gap on the Target population; the corresponding recovery on Test B–Full is 94.3%. Thus, when valid bridge information is limited, APR restrains reconstruction, and when bridge information is sufficiently rich, it nearly reaches the full-information benchmark.

5 Comparative Evaluation under Selective Labels

5.1 Synthetic Benchmark Across Selection Regimes

We compare APR with Unadjusted ERM and five recent approaches for learning under selective or missing labels: DCEM Chang & Wiens (2024), DeepHeckman Cong et al. (2026), SuperDR Zheng et al. (2025), CIC Zheng et al. (2026), and FGD Ha et al. (2024). We consider two complementary selection regimes. *Valid Bridge* retains only a small set of valid Target-like labeled observations, whereas *Unified MNAR* combines outcome-dependent Target selection with a small valid audit channel. Each regime is evaluated under Mild, Moderate, and Severe selection. Core and Target each constitute 50% of the deployment population, but APR is not given this latent composition. Using only observed covariates and label-availability information, the practical Target-share diagnostic in Appendix J yields $\pi_T^{\mathrm{ref}} = 0.50$, which is used by APR as its population mixture weight. Details on auxiliary estimation, outcome-model training, and data handling are provided in Appendix E. Test A contains Target observations only, while Test B–Full contains the full Core–Target deployment mixture. Figure 2 reports performance on both test sets across all selection regimes and severity levels.

Figure 2 reports performance on both test sets across all regimes and severities. APR achieves the highest AUC and lowest mean Log-Loss in every scenario–severity combination. On Test A, its AUC exceeds the strongest non-APR mean by 0.053–0.083 under Unified MNAR and 0.056–0.083 under Valid Bridge. On Test B–Full, the gains are 0.053–0.083 and 0.058–0.084, respectively. Relative to the lowest-Log-Loss non-APR comparator, APR reduces mean Test-A Log-Loss by 0.025–0.048 and Test B–Full Log-Loss by 0.027–0.047. Gains are largest in the Severe Bridge setting, where valid Target evidence is most limited, but APR retains the best mean discrimination and calibration across all regimes. Overall, reconstructing population-relevant risk is most useful when informative Target labels are sparse or selectively observed.

5.2 Evaluation on Bank Marketing Data

We further evaluate APR on the UCI Bank Marketing dataset. We retain the original pre-outcome customer characteristics and campaign-history variables, excluding only the post-contact duration variable, and use them as the empirical basis for a controlled selective-label experiment. Because the original dataset contains no explicit selective-label mechanism, we impose a workflow-motivated selection process in which customer covariates remain available, but campaign outcomes are only partially synchronized across CRM or marketing systems, with a small subset recovered through routine audit or reconciliation.[1] This construction preserves the original feature distribution while creating a reproducible selective-label setting with informative bridges. Using only pre-outcome information, the mechanism assigns about 45% of customers to an under-observed Target population.

[1] This selection design is motivated by prior studies that have shown the selection problem occurs in practice and has adverse consequences Tian et al. (2024); Wernerfelt et al. (2025).

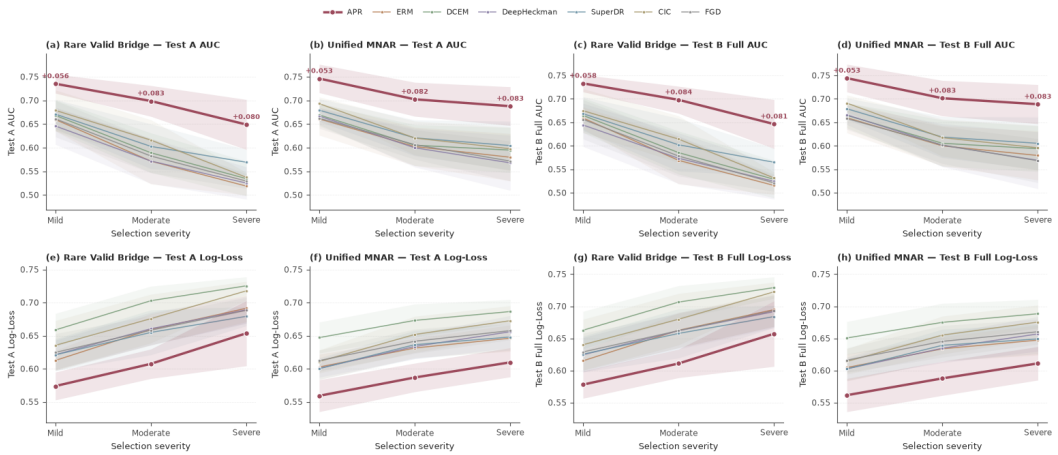


Figure 2: Comparison with existing methods under Valid Bridge and Unified MNAR selection. Panels (a)–(d) report AUC and panels (e)–(h) report Log-Loss on Test A and Test B–Full across three selection severities. Test A contains Target observations only, whereas Test B–Full represents the full Core–Target deployment population. Curves show means across 20 matched random seeds, and shaded regions denote ± 1 standard deviation across seeds. Numbers above the APR curves in the AUC panels report the improvement in mean AUC over the strongest non-APR comparator at the corresponding severity.

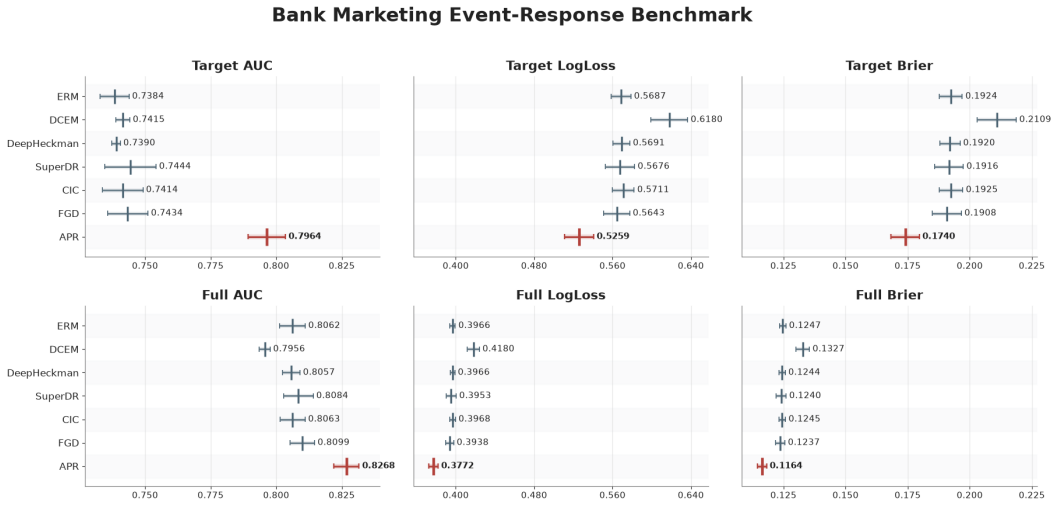


Figure 3: Performance on the Bank Marketing semi-synthetic event-response benchmark. The top row evaluates the under-observed Target population and the bottom row the full population. Points denote means over 20 random seeds and horizontal bars indicate 95% confidence intervals. Higher values are better for AUC, whereas lower values are better for LogLoss and Brier.

In this selected condition, about 95% of Core outcomes but only 4.5% of Target outcomes are observed during training, leaving roughly 590–660 labeled Target cases as bridges. Using observed covariates and label-availability information, the practical Target-share diagnostic in Appendix J yields a reference value of $\pi_T = 0.45$ used by APR without access to latent Core/Target assignments. All methods use identical train/test splits, covariates, and observed labels, with results averaged over 20 random seeds.

As shown in Fig. 3, APR achieves the best mean performance on all six primary metrics, with especially large gains on the under-observed Target population. APR reaches Target AUC 0.7964, compared with 0.7446 for the strongest competitor, SuperDR, while reducing Target LogLoss to 0.5259 and Brier score to 0.1740, versus 0.5643 and 0.1908 for the strongest alternatives. The gains also transfer to the full population: APR obtains Full AUC 0.8268, LogLoss 0.3772, and Brier score 0.1164, compared with best-baseline values of 0.8099, 0.3938, and 0.1237. These results support the intended role of APR in settings where a substantial population component is weakly labeled, but a small number of outcome-reconciled observations retain informative signals about that population.

6 Component Ablation

We next isolate the contribution of APR's main components under the moderate-severity Valid Bridge and Unified MNAR settings, using 30 matched random seeds. We compare full APR with variants that remove selection surprise or Target affinity, replace adaptive global reliability with fixed reliability $\gamma = 0.75$, replace the local gate with a single learned deployment-set constant, and the Fixed Bridge Reweighting baseline, which uses the same bridge-relevance score but does not perform population-risk reconstruction. Figure 7 in Appendix F reports within-seed paired differences in AUC and log-loss, with positive values favoring full APR.

The ablation results show a clear division of labor. Selection surprise provides the largest bridge-identification gain, while Target affinity adds a smaller but stable improvement by screening whether surprising labeled observations are also Target-relevant. Global reliability mainly controls how strongly reconstructed information is trusted, whereas the local gate determines where that information should be transferred. Replacing these adaptive components with fixed counterparts usually hurts performance, especially under Unified MNAR, where the useful degree of adaptation varies across the covariate space.

Full APR also consistently outperforms Fixed Bridge Reweighting. Its Test-A AUC advantages are 0.0228 under Valid Bridge and 0.0187 under Unified MNAR, with corresponding log-loss gains of 0.0152 and 0.0081. Thus APR's gains cannot be explained by bridge-like upweighting alone, and they require combining bridge identification with population-risk reconstruction, reliability control, and local adaptation. The regime contrast is also informative since MNAR amplifies the ranking benefits of adaptive components, while scarce Target supervision can induce larger log-loss penalties in the Valid Bridge setting.

7 Discussion and Limitations

APR is designed for selective-label settings in which the training population is distorted but still contains genuine information about the under-observed group. This occurs when Target outcomes are incompletely observed, yet some Target-relevant labeled observations survive selection and can be overwhelmed by the larger Core sample under ordinary ERM. APR uses these observations to reconstruct the population-level prediction objective rather than assigning outcomes to unlabeled units or simply reweighting the selected sample. By reconstructing separate Core and Target risk components, APR directly addresses distorted population composition when the labeled sample contains too little Target information relative to the deployment population.

The appropriate scope of APR follows from this objective. The Target population must be at least partly identifiable from observed covariates or selection-related information, and bridge outcomes must retain predictive transportability to the under-observed population. When these conditions hold, APR's reliability adjustment and local gating provide an adaptive way to use reconstruction where it is informative and fall back toward ERM where it is not. Performance therefore depends on the quality of the auxiliary selection and Target-affinity models, as well as estimates of Target prevalence. More broadly, APR should be viewed as a population-risk correction layer for selective-label learning, not a standalone architecture. It can be combined with any model supporting observation-weighted empirical risk minimization and shows that selective-label bias need not always be treated as a missing-label problem. When under-observed deployment populations leave valid labeled traces, models can be redirected from performance on observed decisions toward performance on the population they will actually face.

AI USE STATEMENT

The authors used AI-based tools only for writing assistance when editing the paper. These tools were not used for other purposes. All technical content, claims, interpretations, and final text were reviewed and verified by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This research adheres to the ICLR Code of Ethics. All experiments are conducted on publicly available benchmark datasets and environments that do not involve human subjects or personally identifiable information. No conflicts of interest or ethical concerns beyond standard research conduct arise from this work.

REPRODUCIBILITY STATEMENT

We describe the APR procedure, population-risk reconstruction, experimental protocols, evaluation metrics, and main hyperparameters in the paper and appendix. The synthetic data-generating processes and the Bank Marketing semi-synthetic setup are specified in the appendix. Experiments are repeated over multiple random seeds, and we report means and standard deviations where appropriate.

We also include sensitivity analyses for key APR hyperparameters and for misspecification of π_T. Baseline methods are implemented using recommended settings or available implementations, with tuning details described in the appendix. To support reproducibility, we will release code and configuration files needed to reproduce the main results.

REFERENCES

Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134, 2013.

Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: An expectation-maximization framework for learning from biased decisions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6263–6285. PMLR, 2024. URL https://proceedings.mlr.press/v235/chang24b.html.

Lin William Cong, Yanhong Guo, Xin Zhao, and Wenjun Zhou. Deep heckman for loan evaluation. In *NeurIPS 2025 Workshop: Generative AI in Finance*, 2026. URL https://openreview.net/forum?id=SjlIzFqMJD.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. In *Proceedings of the ICML 2018 Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018.

Xiang Gu, Xi Yu, Yan Yang, Jian Sun, and Zongben Xu. Adversarial reweighting for partial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14860–14872, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7ce3284b743aefde80ffd9aec500e085-Abstract.html.

Mingming Ha, Xuewen Tao, Wenfang Lin, Qiongxu Ma, Wujiang Xu, and Linxun Chen. Fine-grained dynamic framework for bias-variance joint optimization on data missing not at random. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104010–104034, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc75254bc4b8b42f401d0ab5d6e9aa4b-Abstract-Conference.html.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. doi: 10.2307/1912352.

Jiayuan Huang, Arthur Gretton, Karsten M. Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, pp. 601–608, 2006.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018. doi: 10.1093/qje/qjx032.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Jun Shao and Lei Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 2016. doi: 10.1093/biomet/asv071.

Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, and Pierre-Alexandre Mattei. Are labels informative in semi-supervised learning? estimating and leveraging the missing-data mechanism. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32521–32539. PMLR, 2023. URL https://proceedings.mlr.press/v202/sportisse23a.html.

Min Tian, Paul R. Hoban, and Neeraj Arora. What cookie-based advertising effectiveness fails to measure. *Marketing Science*, 43(2):407–418, 2024. doi: 10.1287/mksc.2023.1453. URL https://doi.org/10.1287/mksc.2023.1453.

Nils Wernerfelt, Anna Tuchman, Bradley T. Shapiro, and Robert Moakler. Estimating the value of offsite tracking data to advertisers: Evidence from meta. *Marketing Science*, 44(2):268–286, 2025. doi: 10.1287/mksc.2023.0274. URL https://doi.org/10.1287/mksc.2023.0274.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, pp. 903–910. ACM, 2004. doi: 10.1145/1015330.1015425.

Chunyuan Zheng, Haocheng Yang, Haoxuan Li, and Mengyue Yang. Unveiling extraneous sampling bias with data missing-not-at-random. In *Advances in Neural Information Processing Systems*, volume 38, pp. 22512–22541, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1ca17a149e75431e20d5d9878c4f17de-Abstract-Conference.html.

Chunyuan Zheng, Xiang Li, Hang Pan, Eric Wang, Haoxuan Li, Yang Zhang, See-Kiong Ng, and Xiao-Hua Zhou. Causality-based conformal imputation correction with non-random missing labels. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2026. doi: 10.1145/3770855.3818158. URL https://doi.org/10.1145/3770855.3818158.

A Algorithmic Summary of APR

Algorithm 1 summarizes the implementation of Adaptive Population-Risk Reconstruction (APR).

B Additional Methodological and Theoretical Details for APR

This appendix provides additional methodological and theoretical details underlying APR. We first give the detailed specification of the bridge-relevance construction, and then formalize when the observable bridge-relevance score can enrich Target-component labels among selected observations, together with the corresponding non-identifiability boundary. We further give a more explicit justification for interpreting Equation (8) as a reconstructed population-risk objective and characterize the main sources of reconstruction error. Finally, we provide the sharpening of the observation-specific gate and the optional probability-calibration step used in the APR framework.

B.1 Bridge-Score Normalization and Sharpening

We first provide the detailed specification of the normalization, prevalence adjustment, and sharpening transformations used in Eq. (5).

Algorithm 1 Adaptive Population-Risk Reconstruction (APR)

Require: $\mathcal{D} = \{(X_i, G_i, S_i, S_i Y_i)\}_{i=1}^{N}$, Target share π_T, parameters $\alpha, p_T, r_T, p_C, \kappa, \zeta, \tau$
Ensure: Adaptive predictor $\widehat{p}_{\mathrm{APR}}(x)$

/* Stage 1: Bridge discovery */

Fit the auxiliary models on the full sample:

$$\widehat{e}_i = \widehat{P}(S_i = 1 \mid G_i), \qquad \widehat{d}_i = \widehat{P}(S_i = 0 \mid X_i) \tag{4}$$

for each i with $S_i = 1$ **do**
  $b_i \leftarrow \mathcal{N}\Big[(1 - \widehat{e}_i)\widehat{d}_i\Big]$
  $q_i \leftarrow \mathcal{S}_\tau\Big[\alpha b_i + (1 - \alpha)\widetilde{d}_i\Big]$ ▷ Eq. (5)
end for
Freeze $\widehat{e}$, $\widehat{d}$, and $\{q_i\}$

/* Stage 2: Population-risk reconstruction */

for each i with $S_i = 1$ **do**
  $u_i^T \leftarrow q_i^{p_T} \widehat{d}_i^{\,r_T}, \quad u_i^C \leftarrow (1 - q_i)^{p_C}$ ▷ Eq. (7)
end for
$\widehat{f}_0 \leftarrow \arg\min_f \sum_{i:S_i=1} \ell\{f(X_i), Y_i\}$
$\widehat{f}_R \leftarrow \arg\min_f \widehat{R}_{\mathrm{rec}}(f)$ ▷ Eq. (8)

/* Stage 3: Reliability and local adaptation */

Compute ESS_T and c_T from Eq. (11), and set

$$\gamma \leftarrow c_T \left[1 - \exp\left(-\frac{\mathrm{ESS}_T}{\kappa}\right)\right]$$

$\widehat{\widetilde{a}} \leftarrow \arg\min_{\widetilde{a}} \mathcal{L}_{\mathrm{gate}}$ ▷ Eq. (12)
$\widehat{a}(x) \leftarrow \sigma\Big[\zeta \ \mathrm{logit}\{\widehat{\widetilde{a}}(x)\}\Big]$

/* Stage 4: Adaptive prediction */

$\widehat{p}_{\mathrm{APR}}(x) \leftarrow [1 - \gamma\widehat{a}(x)]\,\widehat{p}_0(x) + \gamma\widehat{a}(x)\,\widehat{p}_R(x)$ ▷ Eq. (13)
return $\widehat{p}_{\mathrm{APR}}(x)$

Here, $\mathcal{N}(\cdot)$ is a robust monotone normalization over the labeled sample. For $z_i = (1 - \widehat{e}_i)\widehat{d}_i$, we use

$$\mathcal{N}(z_i) = \mathrm{clip}\left(\frac{z_i - Q^L_{0.50}(z)}{Q^L_{0.995}(z) - Q^L_{0.50}(z) + \epsilon}, 0, 1\right),$$

where $Q^L_\rho(z)$ denotes the empirical ρ-quantile among labeled observations and $\epsilon > 0$ is a small numerical-stability constant.

We define the prevalence-adjusted affinity as

$$\widetilde{d}_i = \mathrm{clip}\left(\widehat{d}_i \frac{\pi_T}{\widehat{\Pr}(S = 0)}, 0, 1\right),$$

where $\widehat{\Pr}(S = 0)$ is the observed unlabeled proportion and π_T is the Target share in the intended deployment population. This adjustment rescales unlabeled affinity to reflect the assumed prevalence of the under-observed Target component. When reliable information about the Target-population share is unavailable, π_T may be treated as a sensitivity parameter; the unadjusted affinity $\widehat{d}_i$ may alternatively replace $\widetilde{d}_i$ to remove its role from bridge scoring, while Eq. (3) still requires a population-mixture value.

The sharpening transformation is

$$\mathcal{S}_\tau(u) = \sigma[\tau \ \mathrm{logit}\{\mathrm{clip}(u, \epsilon, 1 - \epsilon)\}], \qquad \tau > 0,$$

where $\epsilon > 0$ is a small numerical-stability constant. Apart from this endpoint clipping, $\tau = 1$ leaves the input unchanged and larger values produce a sharper relevance score. The parameter $\alpha \in [0, 1]$ controls the relative contribution of direct bridge evidence and population affinity.

B.2 Bridge Enrichment and Identification Boundary

The bridge-relevance construction admits a direct population-level interpretation. The following result formalizes the conditions under which high relevance scores enrich Target-component observations, together with the corresponding observational identification boundary.

Proposition 1 (Bridge enrichment)**.** *Let*

$$e^\star(G) = \Pr(S = 1 \mid G), \qquad d^\star(X) = \Pr(S = 0 \mid X),$$

and let $q^\star$ denote the population analogue of the APR score in Eq. (5)*, obtained by replacing the estimated auxiliary quantities by their population limits. Suppose that, among selected observations,*

$$\big(1 - e^\star(G),\, d^\star(X)\big) \mid (Z = T, S = 1)$$

multivariate first-order stochastically dominates its Core-component counterpart with respect to the coordinatewise order. Then, for every t such that $\Pr(q^\star \geq t \mid S = 1) > 0$,

$$\Pr(Z = T \mid S = 1, q^\star \geq t) \geq \Pr(Z = T \mid S = 1).$$

The inequality is strict whenever

$$\Pr(q^\star \geq t \mid Z = T, S = 1) > \Pr(q^\star \geq t \mid Z = C, S = 1).$$

Thus, high-score labeled observations are enriched in Target-component labels relative to the selected sample.

Corollary 1 (Observational non-identifiability)**.** *If*

$$P(X, G \mid Z = T, S = 1) = P(X, G \mid Z = C, S = 1),$$

then any measurable bridge-discovery score $h(X, G)$ has the same distribution in the two components. Consequently, for every measurable set $\mathcal{A}$ with positive probability,

$$\Pr\{Z = T \mid S = 1, h(X, G) \in \mathcal{A}\} = \Pr(Z = T \mid S = 1).$$

Hence no bridge-discovery rule based only on the observed information can systematically enrich Target labels.

Proposition 1 formalizes how observable Target under-selection and component separation can generate bridge enrichment, whereas Corollary 1 gives the corresponding identification boundary. Appendix C examines both cases empirically. The corresponding proofs follow.

Proof of Proposition 1. Let

$$A^\star = 1 - e^\star(G), \qquad D^\star = d^\star(X).$$

The population bridge score can be written as

$$q^\star = \mathcal{S}_\tau \left[\alpha \mathcal{N}(A^\star D^\star) + (1 - \alpha)\widetilde{D}^\star\right],$$

where $\widetilde{D}^\star$ is the population analogue of the prevalence-adjusted affinity. Because $A^\star, D^\star \in [0, 1]$, $\mathcal{N}(\cdot)$ is monotone, the prevalence adjustment is monotone in $D^\star$, $\alpha \in [0, 1]$, and $\mathcal{S}_\tau(\cdot)$ is increasing for $\tau > 0$, $q^\star$ is coordinatewise nondecreasing in $(A^\star, D^\star)$. Hence the multivariate first-order stochastic dominance assumed in Proposition 1 implies

$$\Pr(q^\star \geq t \mid Z = T, S = 1) \geq \Pr(q^\star \geq t \mid Z = C, S = 1).$$

Let

$$\pi_T^{(1)} = \Pr(Z = T \mid S = 1).$$

By Bayes' rule,

$$\Pr(Z = T \mid S = 1, q^\star \geq t) = \frac{\pi_T^{(1)} \Pr(q^\star \geq t \mid Z = T, S = 1)}{\pi_T^{(1)} \Pr(q^\star \geq t \mid Z = T, S = 1) + (1 - \pi_T^{(1)}) \Pr(q^\star \geq t \mid Z = C, S = 1)}.$$

The stochastic-dominance inequality therefore gives

$$\Pr(Z = T \mid S = 1, q^\star \geq t) \geq \Pr(Z = T \mid S = 1),$$

with strict inequality whenever

$$\Pr(q^\star \geq t \mid Z = T, S = 1) > \Pr(q^\star \geq t \mid Z = C, S = 1).$$

Hence high-score selected observations are enriched in Target-component labels relative to the selected sample. □

Proof of Corollary 1. Suppose that

$$P(X, G \mid Z = T, S = 1) = P(X, G \mid Z = C, S = 1).$$

Then every measurable function $h(X, G)$ has the same conditional distribution in the two components:

$$h(X, G) \mid (Z = T, S = 1) \stackrel{d}{=} h(X, G) \mid (Z = C, S = 1).$$

Consequently, for any measurable set $\mathcal{A}$ with positive probability,

$$\Pr\{h(X, G) \in \mathcal{A} \mid Z = T, S = 1\} = \Pr\{h(X, G) \in \mathcal{A} \mid Z = C, S = 1\}.$$

Applying Bayes' rule gives

$$\Pr\{Z = T \mid S = 1, h(X, G) \in \mathcal{A}\} = \Pr(Z = T \mid S = 1).$$

Thus no bridge-discovery rule based only on the observed information can systematically enrich Target labels. In particular, any such score has chance-level discrimination between Target and Core observations within the selected sample. □

These results formalize the identification logic examined in the diagnostic simulations of Appendix C. Proposition 1 describes the setting in which observable selection and population information can produce Target enrichment, whereas Corollary 1 characterizes the boundary at which such enrichment is impossible.

B.3 Population-Risk Reconstruction

Using the latent component notation introduced in Section 2, let

$$L_i(f) = \ell\{f(X_i), Y_i\},$$

and, for $g \in \{C, T\}$, write the APR component-risk estimator as

$$\widehat{R}_g(f) = \frac{\sum_{i:S_i=1} \widehat{u}_i^g L_i(f)}{\sum_{i:S_i=1} \widehat{u}_i^g},$$

where $\widehat{u}_i^C$ and $\widehat{u}_i^T$ are the soft masses defined in Eq. (7).

For theoretical analysis, let $H = (X, G)$ and let $u_g^\star(H)$ denote a nonnegative reference reconstruction mass. Define

$$R_g^\star(f) := \frac{\mathbb{E}\big[u_g^\star(H) L(f) \mid S = 1\big]}{\mathbb{E}\big[u_g^\star(H) \mid S = 1\big]}.$$

Under exact component identification and outcome transportability, the reference mass satisfies

$$R_g^\star(f) = R_g(f).$$

This equality is the population-level identification condition underlying APR rather than part of the definition of $R_g^\star(f)$. One sufficient interpretation is that: (i) the labeled sample retains support over the population region relevant to component g; (ii) observable covariates and selection-related information are sufficiently informative to recover the component-relevant weighting; and (iii) conditional outcome information carried by valid labeled bridges is transportable to the corresponding deployment component. The negative-control experiment in Appendix C and the transportability stress test directly examine violations of these conditions.

Now suppose that, for some constants $c_g > 0$,

$$\frac{1}{n_1} \sum_{i:S_i=1} \left| \widehat{u}_i^g - c_g u_g^\star(H_i) \right| \{1 + |L_i(f)|\} \xrightarrow{p} 0,$$

where $n_1 = \sum_i S_i$, and suppose that the corresponding oracle denominator is positive and the required expectations are finite. The scale factor c_g cancels because each component risk is normalized. Under the exact identification and transportability condition $R_g^\star(f) = R_g(f)$, the law of large numbers therefore gives

$$\widehat{R}_g(f) \xrightarrow{p} R_g^\star(f) = R_g(f).$$

Consequently, when π_T is correctly specified,

$$\widehat{R}_{\mathrm{rec}}(f) = (1 - \pi_T)\widehat{R}_C(f) + \pi_T \widehat{R}_T(f) \xrightarrow{p} (1 - \pi_T)R_C(f) + \pi_T R_T(f) = R_{\mathrm{pop}}(f).$$

Approximate reconstruction. The same argument yields a useful decomposition when the sufficient conditions hold only approximately. Let $\widehat{R}_g^\star(f)$ denote the empirical risk computed using the oracle masses and let $R_g^\star(f)$ denote its population counterpart. Then

$$\left| \widehat{R}_g(f) - R_g(f) \right| \leq \underbrace{\left| \widehat{R}_g(f) - \widehat{R}_g^\star(f) \right|}_{E_{\mathrm{bridge},g}} + \underbrace{\left| \widehat{R}_g^\star(f) - R_g^\star(f) \right|}_{E_{\mathrm{est},g}} + \underbrace{\left| R_g^\star(f) - R_g(f) \right|}_{E_{\mathrm{transport},g}}.$$

The first term captures imperfect recovery of the component-relevant bridge weighting and therefore includes errors from the auxiliary selection and domain models. The second is finite-sample estimation error and becomes especially important when only a small effective number of bridge observations is available. The third measures failure of bridge outcomes to transport to the corresponding deployment component.

If the reconstruction uses a population share $\widetilde{\pi}_T$ rather than the true π_T, an additional mixture-specification term appears:

$$E_\pi = |\widetilde{\pi}_T - \pi_T| \, |R_T(f) - R_C(f)| \, .$$

Hence the overall discrepancy between reconstructed and population risk can be understood as arising from four sources:

$$\left| \widehat{R}_{\mathrm{rec}}(f) - R_{\mathrm{pop}}(f) \right| \lesssim E_{\mathrm{bridge}} + E_{\mathrm{est}} + E_{\mathrm{transport}} + E_\pi.$$

This decomposition clarifies the empirical design of the paper: Section 4 varies the amount of effective bridge information, the diagnostic simulations examine bridge identification and its failure boundary, the robustness analysis progressively perturbs identification quality and outcome transportability, and the Target-share sensitivity analysis examines E_π.

B.4 Gate Sharpening

The soft population gate in Eq. (12) is trained with class balancing applied when Target-relevant observations are strongly imbalanced. The fitted gate is then sharpened as

$$\widehat{a}(x) = \sigma \left[\zeta \, \mathrm{logit} \left\{ \widehat{\widetilde{a}}(x) \right\} \right], \qquad \zeta > 0.$$

Here, ζ controls the sharpness of local adaptation. Setting $\zeta = 1$ leaves the fitted gate unchanged, while $\zeta > 1$ produces a more decisive separation between Core-like and Target-like observations.

B.5 Estimation and Optional Calibration

The complete procedure follows a modular sequence. First, the selection and domain models are estimated from the combined labeled and unlabeled sample. Their outputs are then frozen and used to construct the continuous relevance scores q_i. Second, the soft Core and Target masses are

calculated, together with ESS_T, c_T, and γ. Third, the base predictor, reconstructed predictor, and soft population gate are estimated using their respective objectives. Prediction for a new observation requires only its task covariates X; the selection indicator, selection variables, and latent population identity are not required at deployment.

For probability-prediction tasks, the two outcome models may operate on somewhat different probability scales. We therefore allow a final monotone calibration step:

$$\widehat{p}(x) = \sigma\left[\frac{\mathrm{logit}\{\widehat{p}^{\,\mathrm{raw}}(x)\}}{T} + c\right], \qquad T > 0. \tag{15}$$

Because $T > 0$, the calibration map is strictly increasing and therefore preserves the ranking of observations and their AUC, while allowing the probability scale to be adjusted for log-loss and Brier score. This calibration step is optional and is not used in the experiments reported in this paper: no post-hoc probability calibration is applied to APR or to any comparator, and all Log-Loss and Brier scores are computed directly from the corresponding uncalibrated model probabilities.

C Diagnostic Simulations for Bridge Identification

We conduct additional controlled simulations to isolate the identification logic underlying bridge relevance, separately from the downstream predictive architecture of Adaptive Population-Risk Reconstruction (APR). The population contains Typical, Target, and irrelevant Outlier observations with shares 0.5, 0.4, and 0.1, respectively. Nonrandom label selection progressively favors the Typical group and underrepresents the Target and Outlier groups, while the marginal label-observation rate is held approximately fixed. This design separates changes in labeled-sample *composition* from changes in the total number of observed outcomes. At an overall observation rate of 0.5, increasing selection severity from $\lambda_{sel} = 0$ to $\lambda_{sel} = 2$ reduces the Target share among labeled observations from approximately 0.40 to 0.16. A size-matched full-label oracle, which uses the same number of labels as ERM but samples them from the full population, improves Target-$S = 0$ AUC relative to ERM by only 0.0030 when $\lambda_{sel} = 0$, but by 0.0197 when $\lambda_{sel} = 2$. Thus, the predictive loss induced by selective labeling is attributable in part to distorted population composition rather than simply to the number of available labels.

We next examine whether target-relevant labeled observations can be identified from the observable selection process. For this diagnostic exercise, we combine cross-fitted *selection surprise*, which is large when an observation is labeled despite having low estimated selection propensity, with an independent measure of *target affinity* to the unlabeled population. This diagnostic score is used only to study the identification principle and should not be interpreted as the final APR estimator. Figure 4 illustrates the complementary roles of the two signals. Under strong selection ($\lambda_{sel} = 2$), average surprise is similar for labeled Target observations (0.654) and irrelevant Outliers (0.667), showing that surprise alone detects selection atypicality but not target relevance. Affinity separates these two groups more clearly (0.652 for Target observations versus 0.262 for Outliers). Their combination yields average diagnostic scores of 0.122, 0.677, and 0.283 for Typical, Target, and Outlier observations, respectively, with Target-versus-Typical and Target-versus-Outlier AUCs of 0.926 and 0.818. Hence, selection surprise and target affinity provide complementary information: the former identifies unexpectedly labeled observations, while the latter helps exclude atypical observations that are unrelated to the underrepresented population.

The identification pattern is not specific to a single selection mechanism. We repeat the experiment using linear, nonlinear, and interaction-based selection functions and vary the information available to the selection model through Clean, Noisy, Partial, and Hidden-group proxy environments. As shown in Figure 5, the diagnostic bridge signal strengthens with selection severity whenever observable information about the latent population structure remains available. Under $\lambda_{sel} = 2$, Target-versus-Typical AUC is approximately 0.63–0.71 across the Clean, Noisy, and Partial environments, while the top 10% of labeled observations ranked by the diagnostic score contain roughly 2.2–2.9 times the baseline Target share. Removing direct group-related information weakens discrimination, but does not eliminate it when target structure remains recoverable from the task covariates.

Finally, we construct a negative-control design that removes the observable distinction between the Typical and Target populations. Let $Z_i \in \{\mathrm{typ}, \mathrm{tar}\}$ denote the latent population group. We impose

$$P(X_i, G_i \mid Z_i = \mathrm{typ}) = P(X_i, G_i \mid Z_i = \mathrm{tar}), \tag{16}$$

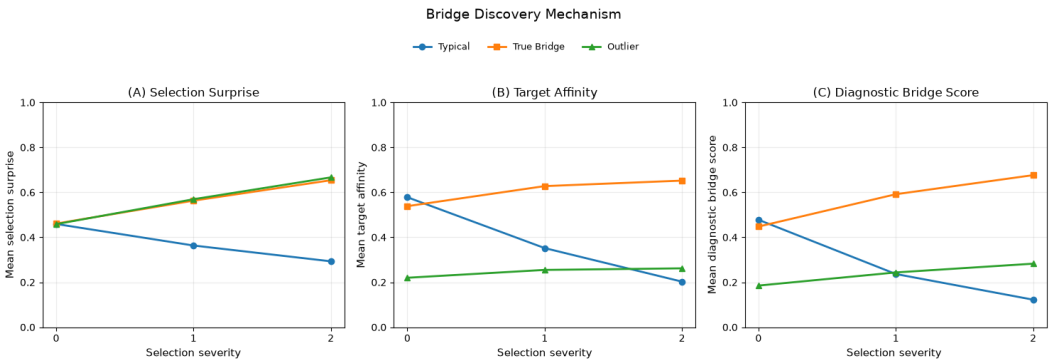


Figure 4: Diagnostic decomposition of bridge relevance at an overall label-observation rate of 0.5. Selection surprise distinguishes unusually labeled observations but does not separate Target observations from irrelevant Outliers. Target affinity supplies the complementary population-relevance signal, and their combination produces substantially stronger discrimination of labeled Target observations.

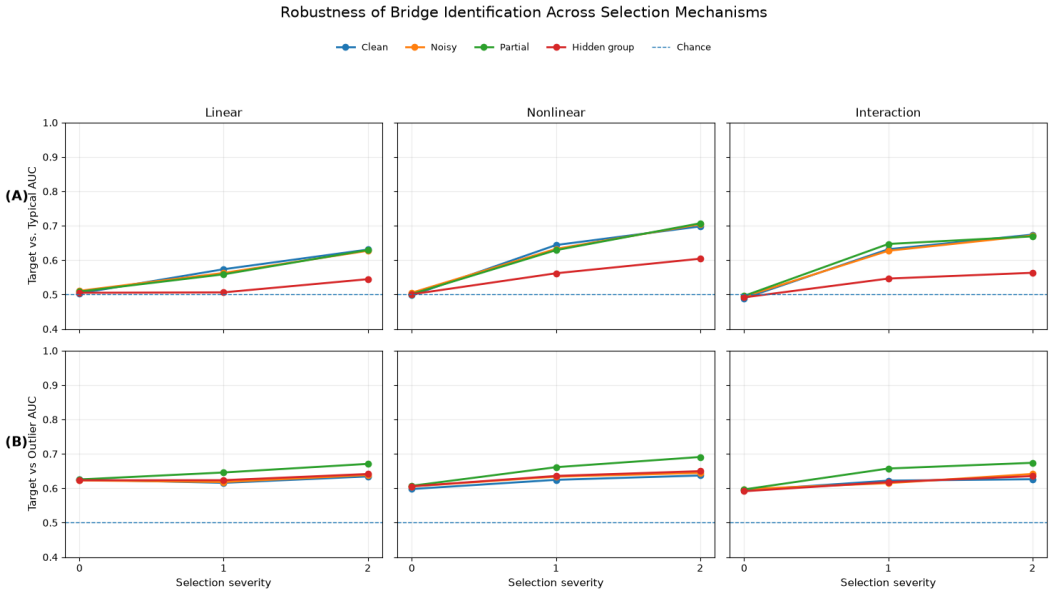


Figure 5: Robustness of bridge identification across linear, nonlinear, and interaction-based selection mechanisms and across different levels of proxy information. The upper row reports Target-versus-Typical discrimination and the lower row reports Target-versus-Outlier discrimination. The dashed line denotes chance-level AUC.

Table 1: Wall-clock training time for ERM and APR across training sample sizes. Values are mean $\pm$ standard deviation over 10 runs.

N_{train}	Hardware	ERM time (s)	APR time (s)	APR / ERM
100	RTX 5070 Laptop GPU	0.492 ± 0.097	2.356 ± 0.212	$4.79\times$
500	RTX 5070 Laptop GPU	0.436 ± 0.048	2.327 ± 0.118	$5.34\times$
1,000	RTX 5070 Laptop GPU	0.426 ± 0.046	2.325 ± 0.244	$5.46\times$
5,000	RTX 5070 Laptop GPU	0.401 ± 0.030	2.156 ± 0.131	$5.37\times$
10,000	RTX 5070 Laptop GPU	0.440 ± 0.075	2.189 ± 0.117	$4.98\times$

while preserving strong selection heterogeneity. Typical and Target observations still have sharply different label-observation probabilities, approximately 0.778 and 0.218, respectively, but their latent identities leave no systematic trace in the observed task covariates and selection proxies available to bridge discovery. As shown in Figure 6, both Selection Surprise and the Diagnostic Bridge Score yield Target-versus-Typical AUCs close to 0.50 across all proxy specifications, while top-decile Target enrichment remains close to one. Thus, strong selection bias alone is insufficient for bridge identification. Some observable information about the latent population structure must remain in the task features, selection proxies, or their representations. This negative control clarifies the identification boundary underlying APR: population reconstruction can exploit informative bridge evidence when such evidence is recoverable, but it cannot manufacture target information when the relevant population distinction is observationally absent.

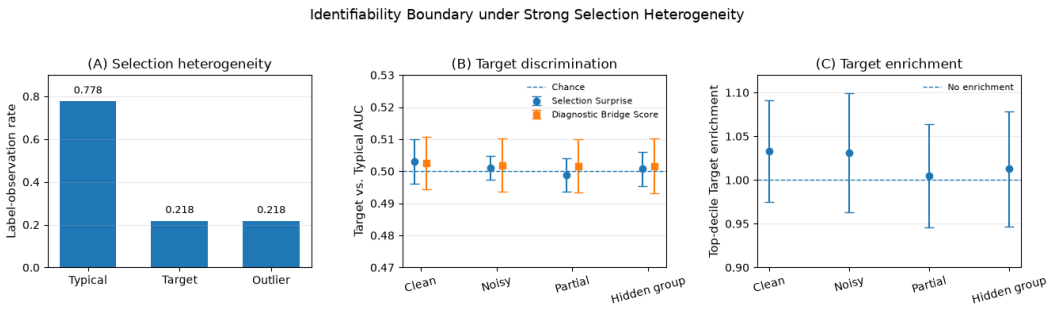


Figure 6: Identifiability boundary under the negative-control design. Although Typical and Target observations retain sharply different label-observation rates (A), their latent population identities are not recoverable from the observed information available to bridge discovery. Consequently, both Selection Surprise and the Diagnostic Bridge Score yield approximately chance-level Target-versus-Typical discrimination (B), while top-decile Target enrichment remains close to one (C). Error bars denote 95% Monte Carlo confidence intervals over 10 replications.

D Computational Cost and Scalability

Table 1 reports wall-clock training time for ERM and APR under the moderate-severity Section 5.1 setting. For each training size, results are averaged over 10 runs across two selection regimes and five random seeds on an NVIDIA GeForce RTX 5070 Laptop GPU. APR time includes the complete training pipeline.

APR requires approximately $5\times$ the wall-clock training time of ERM across the evaluated sample sizes. The relatively stable runtime reflects the fixed-update training protocol, under which all neural networks are trained for fixed numbers of stochastic optimization steps. APR's additional cost mainly comes from the auxiliary bridge-discovery models, population gate, and reconstructed predictor, while bridge scoring and population reconstruction are performed only once.

Table 2: Implementation of the learned components of APR in the experiments.

Object	Model	Training procedure
$\hat{e}_i$	Two-hidden-layer MLP $(64, 32)$ with LayerNorm, SiLU activation, and dropout	Binary classifier for S_i using G_i as input. Predictions are obtained by 2-fold cross-fitting, so each $\hat{e}_i$ is generated out of fold. The model is trained with binary cross-entropy and the sigmoid output estimates $\widehat{\Pr}(S_i = 1 \mid G_i)$.
$\hat{d}_i$	Two-hidden-layer MLP $(64, 32)$ with LayerNorm, SiLU activation, and dropout	Binary classifier for $1-S_i$ using X_i as input. As for $\hat{e}_i$, 2-fold cross-fitting is used to obtain out-of-fold predictions. The sigmoid output estimates $\widehat{\Pr}(S_i = 0 \mid X_i)$.
$\hat{\tilde{a}}(x)$	Two-hidden-layer MLP $(64, 32)$ with LayerNorm, SiLU activation, and dropout	Soft-label classifier trained on labeled observations using q_i as the Target-relevance target. Balanced soft binary cross-entropy gives each observation positive mass proportional to q_i and negative mass proportional to $1 - q_i$.
$\hat{f}_0$	Two-hidden-layer MLP $(64, 32)$ with LayerNorm, SiLU activation, and dropout	Base outcome predictor trained by ordinary binary cross-entropy using all observed outcomes $(S_i = 1)$ with equal observation weights.
$\hat{f}_R$	Same MLP architecture as $\hat{f}_0$	Reconstructed outcome predictor trained on the same observed outcomes as $\hat{f}_0$, but with observation weights induced by the reconstructed Core–Target population risk in Eq. (8).

E Additional Details for Simulation in Section 5

Table 2 summarizes the models used to estimate the main learned quantities in APR.

Training and data-handling details. For APR, the auxiliary models, population gate, base predictor, and reconstructed predictor use the common MLP backbone described in Table 2. They are trained with AdamW using a learning rate of 10^{-3} and weight decay 10^{-5}, with a batch size of 512 and gradient-norm clipping at 5. Task predictors use 240 stochastic updates, whereas auxiliary models use 160 updates. The selection and domain models used to construct $\hat{e}_i$ and $\hat{d}_i$ are estimated by 2-fold out-of-fold cross-fitting; cross-fitting is used for bridge discovery only and is not applied to $\hat{f}_0$ or $\hat{f}_R$. The two outcome predictors use exactly the same network architecture, differing only in their training objectives: $\hat{f}_0$ uses ordinary equal-weight ERM, whereas $\hat{f}_R$ uses the reconstructed population weights from Eq. (8).

All input features are standardized using statistics estimated from the development data only and are then transformed using the same fixed standardization for held-out and test observations. Outcomes with $S_i = 0$ are masked before model fitting and are never provided to APR or any competing estimator. All models use fixed update budgets rather than early stopping.

Section 5.1 uses N_{train}=10,000, $N_{\text{test A}}$=12,000, $N_{\text{test B}}$=18,000. We use relatively large test sets to reduce Monte Carlo evaluation noise in metrics such as AUC and Log-Loss. This helps ensure that variation across random seeds primarily reflects differences arising from the training samples, selection mechanisms, and learning procedures, rather than instability caused by an insufficiently large evaluation set. Meanwhile, Section 5.2 use N_{train}=31,647, $N_{\text{test A}}$=13,564. The training and test sets are randomly split from the full dataset, with almost 70% of observations used for training and almost 30% used for testing.

F Additional Component Ablation Results in Section 6

Figure 7 provides the full component-ablation results.

Table 3: Full results for simulation in Section 5.1 under Valid Bridge and Unified MNAR selection. Values are mean ± standard deviation across random seeds. AUC_A and LL_A denote AUC and Log-Loss on Test A; AUC_B and LL_B denote the corresponding metrics for the Test B. Higher AUC and lower Log-Loss are better. Best results within each scenario–severity block are shown in bold.

		Valid Bridge				Unified MNAR			
Severity	Method	AUC_A	AUC_B	LL_A	LL_B	AUC_A	AUC_B	LL_A	LL_B
Mild	ERM	.659±.024	.659±.023	.613±.015	.615±.016	.661±.034	.660±.032	.603±.019	.605±.021
	DCEM	.668±.032	.665±.037	.659±.025	.663±.029	.669±.023	.665±.027	.647±.023	.651±.025
	DeepHeckman	.646±.039	.644±.046	.621±.022	.624±.027	.667±.018	.666±.020	.601±.013	.603±.016
	SuperDR	.671±.031	.668±.036	.621±.024	.625±.028	.679±.024	.679±.023	.600±.016	.602±.018
	CIC	.679±.043	.675±.046	.635±.037	.640±.039	.693±.014	.691±.017	.611±.017	.614±.020
	FGD	.661±.038	.656±.042	.625±.028	.629±.029	.662±.021	.658±.023	.612±.016	.616±.019
	APR	**.735±.020**	**.733±.018**	**.574±.021**	**.578±.021**	**.746±.030**	**.744±.029**	**.559±.024**	**.561±.026**
Moderate	ERM	.571±.048	.569±.050	.659±.024	.663±.027	.602±.041	.600±.043	.631±.014	.634±.019
	DCEM	.589±.042	.586±.046	.703±.022	.706±.025	.605±.043	.605±.049	.673±.025	.675±.029
	DeepHeckman	.572±.048	.573±.054	.660±.025	.662±.028	.600±.041	.601±.046	.634±.018	.635±.023
	SuperDR	.603±.047	.602±.056	.655±.034	.658±.038	.621±.038	.619±.044	.636±.021	.639±.025
	CIC	.616±.043	.614±.049	.676±.037	.680±.040	.620±.035	.617±.039	.652±.027	.655±.029
	FGD	.582±.036	.578±.043	.658±.022	.662±.027	.606±.031	.602±.039	.641±.018	.645±.025
	APR	**.699±.033**	**.698±.029**	**.607±.022**	**.611±.022**	**.702±.036**	**.701±.038**	**.587±.022**	**.588±.027**
Severe	ERM	.519±.019	.516±.019	.691±.020	.695±.026	.580±.050	.579±.052	.646±.020	.647±.023
	DCEM	.534±.037	.529±.040	.725±.013	.729±.017	.595±.048	.594±.048	.686±.018	.688±.022
	DeepHeckman	.525±.034	.524±.038	.689±.021	.692±.025	.568±.058	.569±.060	.656±.024	.657±.027
	SuperDR	.569±.027	.565±.030	.679±.014	.684±.017	.604±.051	.605±.056	.648±.020	.649±.023
	CIC	.537±.010	.531±.012	.718±.014	.722±.015	.598±.046	.596±.048	.673±.022	.675±.026
	FGD	.529±.015	.521±.018	.688±.019	.693±.025	.572±.042	.568±.047	.658±.014	.660±.019
	APR	**.649±.053**	**.646±.052**	**.654±.049**	**.657±.051**	**.688±.041**	**.689±.042**	**.610±.022**	**.611±.026**

G Robustness to Assumption Violations

We next examine how the proposed APR behaves when its key working assumptions are progressively weakened. We consider three departures from the baseline design. First, *hard-negative contamination* increases the proportion of bridge candidates that resemble the target population in the observed covariates but do not provide valid target information. Second, *identification noise* gradually corrupts the relevance signal used to distinguish informative bridge observations. Third, *transportability violation* makes the outcome relationship learned from the bridge observations increasingly different from that of the target population. For each setting, we conduct 30 paired replications and report the AUC difference between APR and standard ERM. Test A evaluates the Target population, while Test B–Full evaluates the full deployment population containing both Core and Target observations.

Figure 8 shows that APR remains effective under substantial, though not unlimited, departures from the baseline assumptions. Under no violation, APR improves AUC over ERM by approximately 0.096 for the target population and 0.042 for the full population. The method is particularly robust to hard-negative contamination: even when the contamination rate reaches 80%, the corresponding gains remain about 0.075 and 0.032. This indicates that the soft weighting and reliability adjustment prevent misleading bridge candidates from fully dominating the reconstruction.

APR also tolerates considerable identification noise. At a noise intensity of 0.75, it still improves target-population AUC by roughly 0.068. Likewise, under a transport violation of 0.75, APR retains a target-population gain of about 0.034. Its advantage disappears only at the most extreme settings, where the relevance signal becomes entirely uninformative or the bridge outcome mechanism becomes fully nontransportable. Overall, the results show that APR does not depend on perfectly identified bridge observations or exact cross-population invariance. Instead, it continues to deliver meaningful gains whenever the bridge data retain a moderate amount of identification and transport information, while the extreme cases clearly mark the boundary of the method.

H Hyperparameter Sensitivity Analysis

We examine whether the recovery pattern in Section 4 depends on a narrow choice of tuning parameters. We vary one hyperparameter at a time while holding all other components fixed. The six

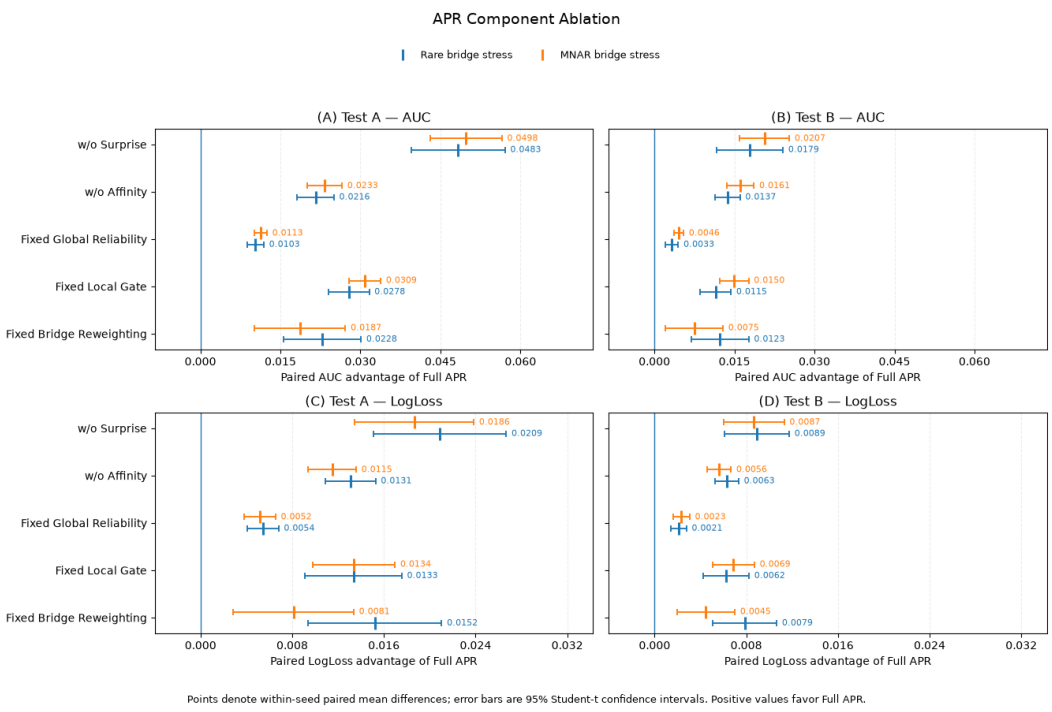


Figure 7: Component ablation of APR under the moderate-severity Section 5.1 settings. Panels report within-seed paired differences in AUC and log-loss for Test A and Test B–Full. Error bars denote 95% Student-t confidence intervals over 30 matched random seeds. Positive values favor full APR.

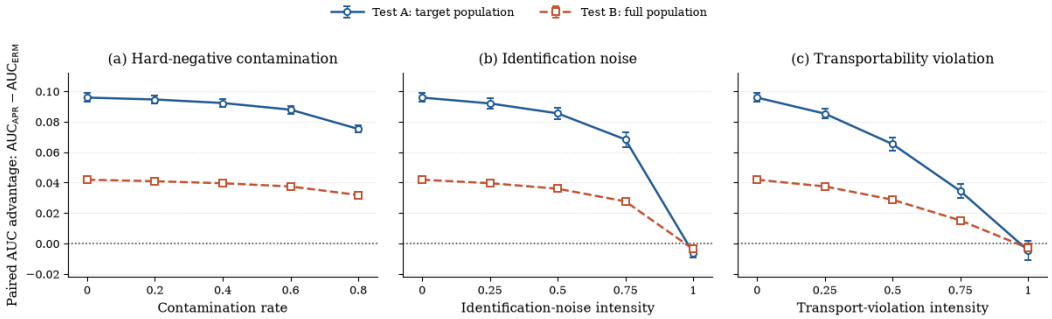


Figure 8: Robustness of APR under progressively weakened assumptions. The vertical axis reports the paired AUC advantage of APR over ERM, and error bars denote 95% Monte Carlo confidence intervals across 30 replications. Positive values favor APR.

parameters are the bridge–affinity mixing weight α, target-relevance power p_T, target-affinity power r_T, ESS reliability scale κ, gate sharpness ζ, and relevance sharpness τ. Their baseline values are

$$(\alpha, p_T, r_T, \kappa, \zeta, \tau) = (0.72, 8, 3, 180, 1.25, 1.50),$$

with sensitivity grids

$$\alpha \in \{0.60, 0.66, 0.72, 0.78, 0.84\}, \quad p_T \in \{4, 6, 8, 10, 12\}, \quad r_T \in \{1, 2, 3, 4, 5\},$$

$$\kappa \in \{100, 140, 180, 240, 300\}, \quad \zeta \in \{1.00, 1.25, 1.50, 2.00\}, \quad \tau \in \{1.00, 1.25, 1.50, 1.75, 2.00\}.$$

The core-mass power p_C is fixed at 2. We evaluate the L2, L4, and L7 regimes from Section 4, corresponding to low, medium, and high bridge information, using 30 Monte Carlo replications.

Figure 9 shows that APR is generally stable over the tested ranges, especially in the medium- and high-information regimes. The effects of α, r_T, and ζ are small. The parameter p_T shows a clearer concentration trade-off, with stronger concentration helping when bridge information is abundant but slightly reducing recovery when information is scarce. The largest low-information sensitivity occurs for κ, where increasing κ makes the reliability adjustment more conservative. Relevance sharpness τ shows a similar but milder information-dependent pattern: stronger sharpening improves recovery in L2, while performance remains close across the tested range in L4 and L7. Overall, the results indicate that APR does not rely on a narrowly tuned parameter configuration and becomes increasingly stable as informative bridge evidence grows.

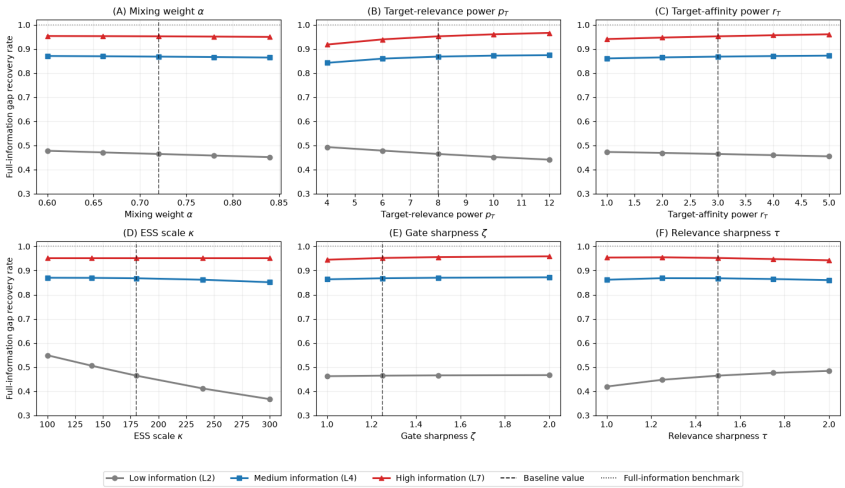


Figure 9: Hyperparameter sensitivity of Adaptive Population-Risk Reconstruction on the Target-only test set. Each panel varies one parameter while holding the others at their baseline values. The vertical dashed line marks the baseline specification and the horizontal dotted line denotes complete recovery of the ERM-to-full-information gap.

I Sensitivity to Target-Share Misspecification

APR uses the assumed Target-population share π_T to connect the bridge-based reconstruction to the intended deployment population. In practice, this population share may only be approximately known. We therefore examine the sensitivity of APR to deliberate misspecification of π_T in the moderate-severity Valid Bridge and Unified MNAR settings from Section 5.1. The true deployment population contains equal proportions of Core and Target observations, so that $\pi_T = 0.50$, while APR is fit using $\widetilde{\pi}_T \in \{0.30, 0.40, 0.50, 0.60, 0.70\}$. The same assumed value is used consistently throughout all APR steps that depend on the Target share, while the data-generating process, train/test distributions, and all other settings are held fixed. Results are based on 20 matched random seeds (seeds 100–119). Figure 10 reports absolute AUC and Log-Loss, while Table 5 reports within-seed paired changes relative to the correctly specified value $\pi_T = 0.50$.

The results show that deployment-population performance is highly stable to Target-share misspecification. For the mixed-population Test B, changing the assumed share from 0.50 to 0.40 or 0.60 changes mean AUC by at most 0.0008 under Valid Bridge and 0.0022 under Unified MNAR, while the corresponding absolute changes in Log-Loss are at most 0.0004 and 0.0013. Even over the wider range from 0.30 to 0.70, the largest Test-B AUC changes are only 0.0013 and 0.0044, respectively. Target-only Test A is more responsive to the assumed population share: underestimating π_T reduces Target AUC, whereas increasing $\widetilde{\pi}_T$ places greater emphasis on the reconstructed Target component and improves Target discrimination in these finite-sample settings. This behavior should not be interpreted as evidence that the true population share is incorrectly specified. The true π_T defines the deployment-population mixture in the population-risk estimand, whereas the value maximizing a finite-sample predictive metric, particularly on a Target-only test set, need not coincide with that population quantity. Overall, the experiment indicates that APR does not require π_T to

Table 4: Hyperparameter sensitivity of the full-information gap recovery rate. "Baseline" reports recovery under the main Section 4 specification, while "Range" gives the minimum and maximum recovery over the corresponding sensitivity grid.

		Test Set A		Test Set B	
Parameter	Regime	Baseline	Range	Baseline	Range
α	Low	0.4643	[0.4513, 0.4777]	0.4279	[0.4170, 0.4392]
	Medium	0.8679	[0.8641, 0.8705]	0.8454	[0.8428, 0.8469]
	High	0.9520	[0.9496, 0.9533]	0.9429	[0.9418, 0.9431]
p_T	Low	0.4643	[0.4409, 0.4931]	0.4279	[0.4052, 0.4576]
	Medium	0.8679	[0.8421, 0.8739]	0.8454	[0.8246, 0.8482]
	High	0.9520	[0.9179, 0.9663]	0.9429	[0.9103, 0.9563]
r_T	Low	0.4643	[0.4546, 0.4728]	0.4279	[0.4186, 0.4363]
	Medium	0.8679	[0.8604, 0.8720]	0.8454	[0.8398, 0.8477]
	High	0.9520	[0.9407, 0.9605]	0.9429	[0.9322, 0.9508]
κ	Low	0.4643	[0.3670, 0.5489]	0.4279	[0.3484, 0.4830]
	Medium	0.8679	[0.8515, 0.8700]	0.8454	[0.8308, 0.8472]
	High	0.9520	[0.9520, 0.9520]	0.9429	[0.9429, 0.9429]
ζ	Low	0.4643	[0.4620, 0.4664]	0.4279	[0.4258, 0.4292]
	Medium	0.8679	[0.8635, 0.8718]	0.8454	[0.8411, 0.8482]
	High	0.9520	[0.9448, 0.9589]	0.9429	[0.9348, 0.9499]
τ	Low	0.4643	[0.4191, 0.4840]	0.4279	[0.3811, 0.4499]
	Medium	0.8679	[0.8603, 0.8683]	0.8454	[0.8255, 0.8461]
	High	0.9520	[0.9424, 0.9549]	0.9429	[0.9318, 0.9429]

Table 5: Paired sensitivity of APR to Target-share misspecification. Entries report the mean within-seed change relative to the correctly specified $\widetilde{\pi}_T = 0.50$, with 95% Student-t confidence intervals over 20 matched seeds. For AUC, positive values indicate higher AUC than under the correctly specified share; for Log-Loss, positive values indicate higher (worse) Log-Loss.

Scenario	$\widetilde{\pi}_T$	Δ Test A AUC	Δ Test B AUC	Δ Test A Log-Loss	Δ Test B Log-Loss
Valid Bridge	0.30	$-0.0097\,[-0.0132, -0.0061]$	$+0.0013\,[-0.0014, +0.0041]$	$+0.0027\,[-0.0018, +0.0071]$	$+0.0004\,[-0.0018, +0.0025]$
	0.40	$-0.0044\,[-0.0061, -0.0028]$	$+0.0001\,[-0.0011, +0.0013]$	$+0.0019\,[-0.0006, +0.0045]$	$+0.0004\,[-0.0007, +0.0016]$
	0.60	$+0.0042\,[+0.0028, +0.0056]$	$+0.0008\,[-0.0001, +0.0017]$	$-0.0001\,[-0.0042, +0.0040]$	$+0.0002\,[-0.0017, +0.0021]$
	0.70	$+0.0063\,[+0.0040, +0.0085]$	$+0.0013\,[-0.0002, +0.0027]$	$-0.0050\,[-0.0074, -0.0025]$	$-0.0015\,[-0.0028, -0.0003]$
Unified MNAR	0.30	$-0.0104\,[-0.0132, -0.0077]$	$-0.0009\,[-0.0026, +0.0008]$	$+0.0063\,[+0.0025, +0.0101]$	$+0.0016\,[-0.0001, +0.0033]$
	0.40	$-0.0051\,[-0.0064, -0.0039]$	$-0.0009\,[-0.0017, -0.0001]$	$+0.0034\,[+0.0009, +0.0059]$	$+0.0009\,[-0.0003, +0.0020]$
	0.60	$+0.0055\,[+0.0043, +0.0068]$	$+0.0022\,[+0.0014, +0.0030]$	$-0.0044\,[-0.0063, -0.0024]$	$-0.0013\,[-0.0021, -0.0005]$
	0.70	$+0.0100\,[+0.0079, +0.0121]$	$+0.0044\,[+0.0031, +0.0057]$	$-0.0080\,[-0.0111, -0.0050]$	$-0.0024\,[-0.0037, -0.0011]$

be known precisely: a reasonably informed population-share estimate is sufficient to preserve the deployment-level benefit of the reconstruction.

J Practical Specification of the Target-Population Share

In applications, the Target component may be latent even when its approximate population share is known from external sources. When reliable deployment-level information is available, such as administrative records, system-level population statistics, or prior operational knowledge, we recommend specifying π_T directly from that information. When no credible external estimate is available, a simple observable diagnostic can instead be used to select a coarse reference value. Recall that

$$\hat{d}_i = \widehat{\Pr}(S_i = 0 \mid X_i)$$

measures affinity with regions in which labels are typically unobserved. We therefore compute the population fraction

$$\hat{r}_T = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{d}_i \geq 0.5\},$$

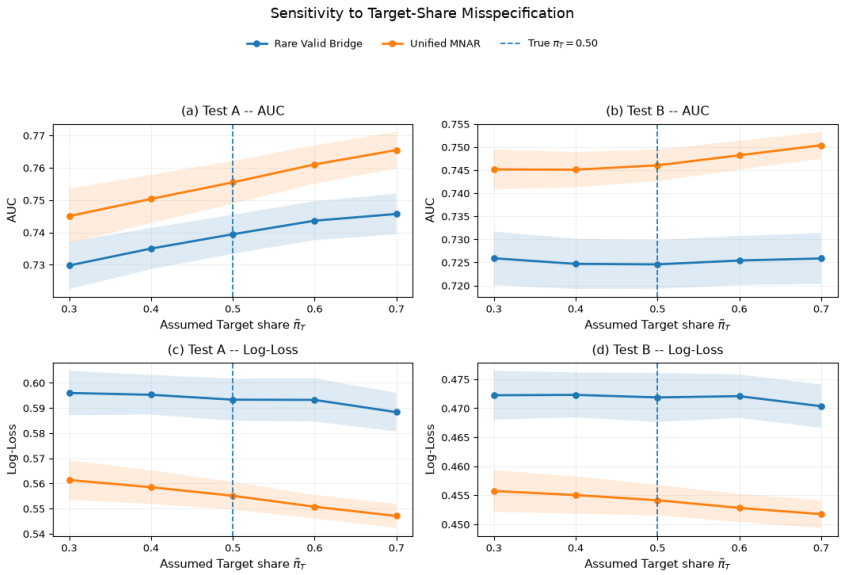


Figure 10: Sensitivity of APR to misspecification of the Target-population share. The true deployment share is $\pi_T = 0.50$ (vertical dashed line), while APR is fit using $\widetilde{\pi}_T \in \{0.30, 0.40, 0.50, 0.60, 0.70\}$. Panels (a) and (b) report AUC on the Target-only Test A and the mixed deployment-population Test B, respectively; panels (c) and (d) report the corresponding Log-Loss. Curves show means over 20 matched random seeds and shaded regions denote 95% Student-t Monte Carlo confidence intervals. Test-B performance remains nearly unchanged over a broad range of assumed Target shares, whereas Test-A performance responds more systematically to the strength placed on the reconstructed Target component.

and round $\hat{r}_T$ to the nearest 0.05 to obtain a practical reference value π_T^{ref}. The threshold 0.50 has the natural interpretation that an observation is predicted to be more likely unlabeled than labeled. Importantly, $\hat{r}_T$ is not treated as an estimator of the latent Target prevalence, since $\hat{d}_i$ is a measure of under-observed-region affinity rather than $\Pr(Z_i = T \mid X_i)$. Its purpose is only to identify the approximate scale of the under-observed population component.

Figure 11 illustrates this diagnostic using the moderate-severity settings from experiment in Section 5.1 while withholding the true Target identities and true π_T from the procedure. The empirical affinity distributions exhibit a clear separation between well-observed and strongly under-observed regions. The fractions with $\hat{d}_i \geq 0.5$ are 0.519 ± 0.003 under Valid Bridge and 0.522 ± 0.004 under Unified MNAR, leading in both cases to $\pi_T^{\text{ref}} = 0.50$. The true population share, revealed only after this diagnostic, is $\pi_T = 0.50$ in both settings. Thus, in this controlled example, the observable affinity structure is sufficient to recover the correct practical reference scale without access to latent population membership.

When the affinity distribution does not exhibit a sufficiently clear under-observed region, this diagnostic should not be interpreted as identifying the Target share. In such cases, π_T is better treated as a sensitivity parameter over a plausible range. Appendix I shows that APR's deployment-population performance remains stable under moderate misspecification of this quantity.

K Hyperparameter Choices

Table 6 summarizes the structural hyperparameters used by APR in the comparative simulation experiments. These parameters control the relative contribution of bridge evidence and Target affinity, the concentration of the reconstructed component masses, and the strength of the global and local adaptation mechanisms.

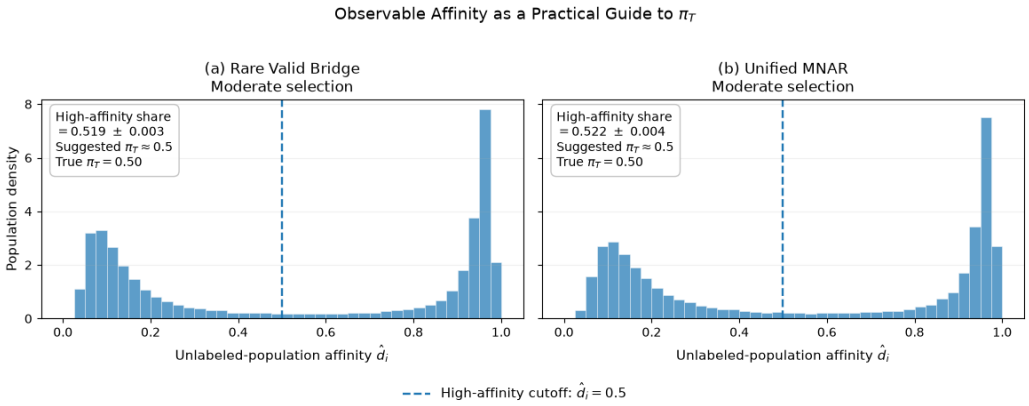


Figure 11: Observable-affinity diagnostic for selecting a practical reference value of π_T. The vertical dashed line marks $\hat{d}_i = 0.5$. The population mass to the right of this threshold provides a coarse indication of the size of the strongly under-observed region. The true $\pi_T = 0.50$ is used only for post-hoc validation and is not supplied to the diagnostic procedure.

Table 6: Baseline hyperparameter configuration for APR.

Parameter	Baseline value	Role
α	0.72	Weight on direct bridge evidence relative to Target affinity
p_T	8	Concentration of Target mass on high-relevance observations
r_T	3	Additional concentration according to Target affinity
p_C	2	Concentration of Core mass as Target relevance decreases
κ	180	Scale controlling the growth of global reconstruction reliability
ζ	1.25	Sharpness of the observation-specific population gate
τ	1.50	Sharpness of the continuous bridge-relevance score

The baseline APR configuration was selected once using an independent development sample of $N_{\text{val}} = 1{,}000$. For each tunable structural hyperparameter, we prespecified a small grid of four or five plausible values and searched over the resulting candidate configurations. The configuration with the highest development-set AUC was selected. The resulting configuration was then fixed across the Valid Bridge and Unified MNAR settings, all three selection-severity levels, and all evaluation seeds; no scenario- or test-seed-specific retuning was performed, and the evaluation test sets were not used for hyperparameter selection. For APR and all competing methods that require hyperparameter tuning, model selection is conducted on the same held-out development sample, using AUC as the common selection criterion. Importantly, the exact choice of these baseline values is not critical to the reported conclusions. Appendix H varies the main APR hyperparameters around this specification and shows that performance is generally stable under moderate perturbations, particularly when bridge information is sufficiently informative. Thus, the reported gains do not rely on a narrowly tuned hyperparameter configuration.

For all competing methods, we follow the hyperparameter specifications provided in the original papers or the authors' released implementations. When explicit default or recommended values are available, we retain those settings. When the original work reports a hyperparameter search grid, we search over the reported grid and select the configuration according to performance on the held-out development set. The evaluation test sets are not used for hyperparameter selection. This protocol is applied consistently across DCEM Chang & Wiens (2024), DeepHeckman Cong et al. (2026), SuperDR Zheng et al. (2025), CIC Zheng et al. (2026), and FGD Ha et al. (2024).