DETECTING HIDDEN CONFORMAL FAILURES UNDER FEATURE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Split conformal prediction provides marginal coverage under exchangeability, yet deployment-time feature shift can silently invalidate this guarantee even when point accuracy barely changes. We study label-free detection of such hidden coverage failures. Soft Conformal Boundary Sensitivity (SCBS) profiles how source-side feature interventions move examples across the conformal boundary and combines this vulnerability with null-calibrated unlabeled target drift. Across 57,600 controlled scenarios, SCBS detects five-point hidden coverage losses with dataset-macro AUROC 0.846. Direction-aware source simulation and an output-based monitor expose complementary mechanisms, so we select between them using a source-only router. Frozen on an ordered Bank Marketing audit, the router obtains AUROC/AUPRC $0.974 \pm 0.037/0.997 \pm 0.003$; across three natural hidden-failure streams it obtains $0.965 \pm 0.025/0.894 \pm 0.035$ with FPR 0.146. Strong generic alternatives do not isolate harmful drift: classifier two-sample testing and linear-time MMD obtain AUROC 0.582 and 0.746 and alarm on nearly every benign shifted window. An impossibility result formalizes why unlabeled covariates cannot universally certify target coverage, and a finite-label extension restores design-based identification. The result is a mechanism-aware deployment warning, not an unconditional coverage guarantee.

1 INTRODUCTION

A conformal predictor can fail before its classifier does. Split conformal prediction turns the scores of an arbitrary model into prediction sets with finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Romano et al., 2020; Angelopoulos et al., 2021). After deployment, however, a shifted population may preserve the classifier's argmax on most examples while moving true-label scores across the calibrated set boundary. Point accuracy then appears stable even as coverage falls. We call this operationally important mismatch a *hidden conformal failure*.

Generic shift detection does not resolve the problem. A two-sample test asks whether source and target covariates differ, not whether the difference moves probability mass across the fixed conformal threshold. Missing columns, rescaled sensors, altered dependencies, and benign seasonal changes can all make the domains easily distinguishable (Ovadia et al., 2019; Cheng et al., 2025; Peng et al., 2025). In a persistent stream, an omnibus detector may therefore alarm on almost every later window even though only a minority suffer material undercoverage. The deployment question is more selective: *which observable changes threaten coverage while ordinary accuracy remains nearly unchanged?*

We study a classifier and split-conformal threshold that are frozen before deployment. Labeled source calibration data remain available, whereas each target window contains covariates and model outputs but no labels. The goal is to rank or alarm on windows whose coverage has dropped by an operator-chosen tolerance γ, emphasizing five-point coverage losses with at most a one-point accuracy drop. Unlike work that adapts prediction sets under a specified shift model (Tibshirani et al., 2019; Gibbs & Candès, 2021; Barber et al., 2023), we monitor the failure of an already-deployed set without assuming an accurate source-to-target density ratio.

This problem has an unavoidable information limit. The same target covariate marginal can be paired with two conditional label laws that place labels inside versus outside the fixed prediction set. Their coverages can differ by the mass of a positive-measure region, yet every label-free monitor observes the same data. Section 2 formalizes this non-identifiability. It rules out a universal label-free

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.

certificate and changes the scientific question: rather than estimating target coverage from unlabeled data without assumptions, what source-side structure can make a coverage-harming shift detectable?

We propose the *boundary-alignment principle*: a target shift is evidence of coverage risk only when it occurs along directions to which the fixed conformal boundary is vulnerable. Soft Conformal Boundary Sensitivity (SCBS) operationalizes this principle. Offline source interventions measure how strongly each feature direction moves examples across the calibrated threshold. Online, null-calibrated unlabeled statistics measure where the target window moved. Their alignment, rather than shift magnitude alone, produces the risk score. Consequently SCBS estimates neither generic feature importance nor generic domain discrepancy; it targets *boundary-relevant shift*.

The theory makes the structural bet explicit. Under probe-alignment and drift-proxy conditions, target coverage loss decomposes into the SCBS score plus four interpretable terms: probe mismatch, drift-proxy error, unrepresented shift, and soft-boundary approximation. A large score is thus evidence under alignment, while a small score is reassuring only when these remainder terms are controlled. Direction-aware interventions, source-simulated calibration, and routing to a complementary output channel are deployment refinements of this core principle. If a small uniform sample of target labels is available, a design-based probe changes the information setting and restores identification.

The experiments test the principle rather than only compare systems. Across 57,600 controlled feature-shift scenarios, SCBS obtains dataset-macro AUROC 0.846 for hidden five-point failures and improves over mean-score sensitivity, generic drift, and predictive-importance-weighted drift. It retains its ordering on held-out datasets, corruption families, and severities, directly stress-testing probe alignment. On three natural hidden-failure streams, classifier two-sample testing and MMD detect pervasive domain change but alarm on nearly every benign shifted window; the boundary-aware channel is substantially more selective, and source-only routing reaches AUROC/AUPRC 0.965 ± 0.025/0.894 ± 0.035 with FPR 0.146. Negative and accuracy-visible audits expose where the structural assumption does not transfer.

Our contributions are threefold:

- We formulate hidden conformal failure as coverage degradation of a fixed set that is largely invisible to point accuracy, and establish its non-identifiability from source data and unlabeled target covariates under arbitrary joint shift.
- Motivated by this limit, we introduce the boundary-alignment principle and SCBS. Its conditional surrogate analysis separates four explicit ways in which source probes and target mechanisms can disagree.
- We evaluate this principle across controlled, held-out, external, and natural shifts, including regimes where generic detectors saturate. Source-only deployment refinements and a uniform-label probe respectively improve operation under structure and close the identification loop when labels become available.

2 Problem Formulation

Let P_S denote the source distribution and let a classifier fixed before deployment output $p_\theta(y \mid x)$. We use the nonconformity score $s(x, y) = 1 - p_\theta(y \mid x)$. From n_{cal} labeled source calibration examples, split conformal chooses the usual order-statistic threshold q at error level α and returns $\widehat{C}(x) = \{y : s(x, y) \le q\}$. The classifier, score function, calibration sample, and q remain frozen thereafter. Deployment supplies a sequence of unlabeled target windows $X_{1:n}^T \sim P_{T,X}$ whose feature marginals or dependencies may differ from the source.

The operational estimand is coverage of this already-deployed set, $C(P) = \Pr_{(X,Y)\sim P}\{Y \in \widehat{C}(X)\}$. Let C_S be coverage on an independent clean-source evaluation distribution and $C_T = C(P_T)$. We define the nonnegative degradation

$$G(P_T) = [C_S - C_T]_+,$$

and ask a monitor to rank windows, or to alarm, when $G(P_T) \ge \gamma$ for an operator-chosen tolerance $\gamma \in \{0.05, 0.10\}$. The difficult endpoint is a *hidden failure*: $G(P_T) \ge \gamma$ while the target point-accuracy drop is at most 0.01. This definition excludes ordinary accuracy collapse, for which existing output or performance monitors already provide strong signals.

At monitoring time the label-free detector observes labeled source data, the fixed model and conformal rule, and $X^T_{1:n}$ together with any model outputs computable from it. It does not observe target labels. Those labels are revealed only after scoring to construct retrospective endpoints and metrics. Statistical replication is at the dataset, predictor-seed, and non-overlapping-window levels; multiple perturbations or correlated windows are not treated as independent evidence. Section 6 therefore reports cell-wise metrics followed by dataset-macro aggregation and paired dataset or moving-block uncertainty intervals.

2.1 THE INFORMATION LIMIT

Before constructing a monitor, we ask whether the observable information can identify the target coverage at all.

Proposition 1 (Unlabeled non-identifiability; informal)**.** *Fix a nontrivial set predictor $\widehat{C}$ and a target covariate marginal Q_X. If a measurable set A of positive Q_X-mass satisfies $\emptyset \neq \widehat{C}(x) \neq \mathcal{Y}$, there exist two target joint distributions with the same Q_X and identical observable unlabeled data but whose coverages differ by $Q_X(A)$.*

One construction assigns labels inside $\widehat{C}(x)$ on A and the other assigns them outside. Every method observing only source data and unlabeled target covariates consequently has the same output distribution in both worlds. No such method can universally certify target coverage; Appendix A.2 gives the formal statement and proof.

This negative result changes the goal. We do not seek a distribution-free coverage estimator from unlabeled covariates. Instead, we ask which source-side structural evidence can turn an observable target shift into a useful warning. The next section introduces one such inductive bias: alignment between where the target moved and where the fixed conformal boundary is vulnerable.

We also study a separate, explicitly semi-supervised decision problem in which a uniform sample of m target labels may be queried. Keeping this resource boundary visible is essential: unlabeled monitoring can prioritize an audit, whereas even a small probability sample of labels identifies target coverage under a design-based argument.

3 BOUNDARY-ALIGNED MONITORING

The impossibility result rules out inference from target movement alone. We therefore introduce the *boundary-alignment principle*: observable shift is evidence of coverage risk only when it occurs along directions that move source probability mass across the fixed conformal threshold. Figure 1 shows the distinction. Labeled source data reveal the boundary-vulnerable directions; an unlabeled target window reveals where deployment moved. Their alignment produces a warning score but neither estimates target coverage without assumptions nor alters the conformal set.

3.1 FROM THE PRINCIPLE TO SCBS

For temperature $\tau > 0$, replace the hard boundary indicator by

$$\phi_\tau(s; q) = \sigma((s - q)/\tau).$$

For feature j, independently permute its source calibration values and define the positive soft-miscoverage increase

$$v_j = \left[\mathbb{E}\{\phi_\tau(S^{\pi_j}; q) - \phi_\tau(S; q)\}\right]_+ .$$

This quantity is large when breaking feature j's source relation pushes probability mass across the conformal boundary. It differs from mean-score sensitivity because it emphasizes scores close to q.

From an unlabeled target batch, we compute feature-wise marginal and dependency drift. Each raw statistic is standardized against bootstrap source-vs-source comparisons, yielding a nonnegative null-calibrated score d_j. The proposed deployment risk is

$$R_{\mathrm{SCBS}} = \sum_{j=1}^{d} v_j d_j.$$

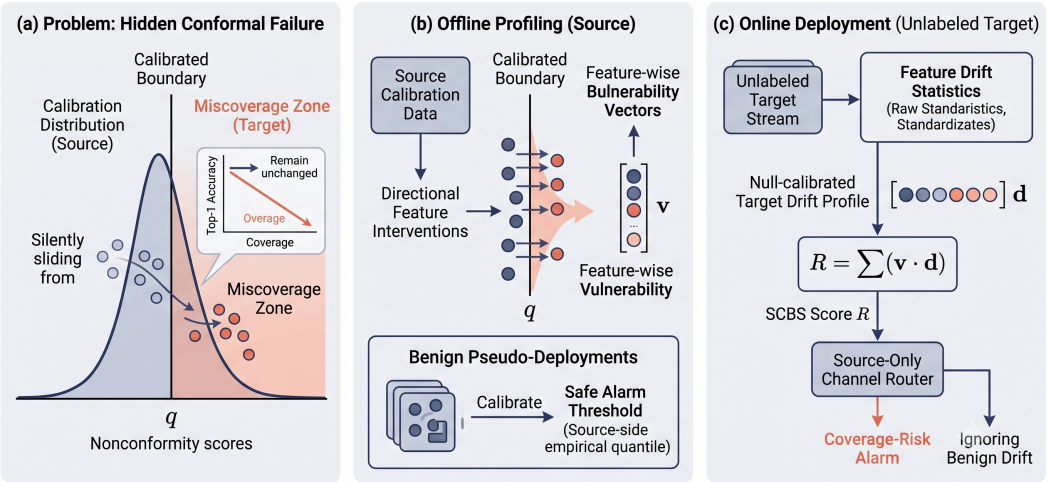


Figure 1: Boundary-aligned monitoring. A hidden failure moves true-label scores across the fixed threshold without necessarily changing the argmax. Source interventions identify vulnerable directions, while an unlabeled target window identifies where deployment moved. SCBS warns when these two pieces of evidence align; source calibration and routing are operating refinements rather than coverage guarantees.

The source profile v is computed offline for each model and calibration split. At deployment, only the target drift profile and a dot product are required.

The product encodes a deliberately asymmetric hypothesis. Large drift in a feature with $v_j \approx 0$ is observable but not locally connected to miscoverage; large vulnerability without target movement is a dormant risk. Only their co-occurrence raises the score. Null standardization also makes heterogeneous marginal and dependency statistics comparable across features and prevents naturally noisy coordinates from dominating solely because of their scale. Neither factor alone identifies the target conditional label law, which is why Section 4 treats SCBS as a structured surrogate rather than a certificate.

This estimand differs from predictive importance. Permutation importance asks whether a feature affects predictions or average loss; v_j asks whether movement along that coordinate pushes probability mass through the particular threshold q. SCBS is therefore a diagnostic of boundary-relevant shift, not a generic feature-importance score.

3.2 Deployment Refinements

Direction-aware score. Magnitude-only drift treats movements toward safer and more failure-prone regions identically. We therefore evaluate two local source interventions $X_j^+ = X_j + h$ and $X_j^- = X_j - h$, where $h = 0.5$ after source standardization, and form separate boundary vulnerabilities

$$v_j^+ = \left[\mathbb{E}\{\phi_\tau(S(X_j^+); q) - \phi_\tau(S(X); q)\}\right]_+, \tag{1}$$

$$v_j^- = \left[\mathbb{E}\{\phi_\tau(S(X_j^-); q) - \phi_\tau(S(X); q)\}\right]_+. \tag{2}$$

For an unlabeled target batch, its signed standardized mean displacement is itself standardized by source bootstrap resampling, producing $z_j \in [-20, 20]$. The directional marginal score is

$$R_{\text{dir,marg}} = \sum_j v_j^+ [z_j]_+ + v_j^- [-z_j]_+.$$

Dependency changes do not have an intrinsic positive or negative direction, so we retain the original permutation vulnerability for the null-calibrated dependency component d_j^{dep}. The final ranking score is

$$R_{\text{dir}} = R_{\text{dir,marg}} + \sum_j v_j d_j^{\text{dep}}.$$

All vulnerabilities and bootstrap scales use only the labeled source calibration fold.

Source-simulated operating point. A source-vs-source null controls ordinary sampling variation but can saturate after a persistent benign domain change. To set a more deployment-relevant operating point, we construct a fixed dictionary of source pseudo-deployments: feature subsets are selected from high-importance, low-importance, and random features; missingness, permutation, noise, and rescaling interventions are applied at four prespecified severities. For each pseudo-deployment we can measure its coverage loss because source labels are available. The alarm threshold is the empirical 95th percentile of R_{dir} among source pseudo-deployments whose coverage loss is below five percentage points. Unmodified finite-sample controls are included in the dictionary. We call the resulting alarm *Dir-SimCal*. Natural target labels are used only for final evaluation, never for score construction or threshold selection.

Source-only routing. Feature-side boundary movement and output-side suitability can fail in different regimes. Let R_{dir} be the Dir-SimCal score and let R_{out} be Output-UPE. Within each model–nominal-level–window-size cell, we normalize each channel by its 95th percentile over benign source pseudo-deployments. When at least five harmful pseudo-deployments are available, we select the channel with the larger harmful-versus-benign empirical AUROC; otherwise we fall back to Output-UPE. The selected normalized score and its unit threshold are frozen for deployment. This router uses labeled source simulations but neither labels nor outcomes from a natural target stream. Appendix A.7 gives a finite-sample wrong-selection bound under an iid pseudo-deployment model.

Optional labels close the identification loop. Arbitrary changes in $P(Y \mid X)$ cannot generally be identified from unlabeled target covariates. When a small target-label budget is available, we therefore augment the label-free alarm with a design-based probe. Let $M_i = \mathbf{1}\{Y_i \notin \widehat{C}(X_i)\}$ for a target window of size n, let $\widehat{m}_i$ be a source-trained prediction of M_i from covariates and model-output signals, and let L be a uniform sample of m target indices. If $\overline{M}_S$ is source miscoverage, define

$$\widehat{G}_{\mathrm{naive}} = m^{-1} \sum_{i \in L} M_i - \overline{M}_S, \tag{3}$$

$$\widehat{G}_{\mathrm{MA}} = n^{-1} \sum_{i=1}^{n} \widehat{m}_i + m^{-1} \sum_{i \in L} (M_i - \widehat{m}_i) - \overline{M}_S. \tag{4}$$

The second expression is a finite-population, model-assisted regression estimator (Breidt & Opsomer, 2017). Under uniform sampling it is design-unbiased for the target-source miss-rate gap, even if the source-trained predictor is misspecified. Its standard error is the labeled-residual standard deviation divided by $\sqrt{m}$, multiplied by $\sqrt{(n-m)/(n-1)}$. We distinguish a screening alarm $\widehat{G} > z_{0.95}\widehat{\mathrm{se}}$ from stricter certification $\widehat{G} - z_{0.95}\widehat{\mathrm{se}} > \gamma$. The queried labels are thus an explicit additional resource and are never attributed to Dir-SimCal.

4 When Boundary Alignment Is Informative

Section 2 establishes that unlabeled target covariates cannot universally identify coverage. We now state what the boundary-alignment assumption buys. The result is conditional: it formalizes when SCBS is a diagnostic surrogate and exposes the terms that make the surrogate fail.

Write $M(P; q) = \Pr_P(S > q)$ and $M_\tau(P; q) = \mathbb{E}_P \phi_\tau(S; q)$. Soft and hard miscoverage differ only by score mass near the boundary and a temperature term: for every $\epsilon > 0$,

$$|M_\tau(P; q) - M(P; q)| \le \Pr_P(|S - q| \le \epsilon) + e^{-\epsilon/\tau}.$$

This inequality explains the boundary-local probe: when little probability mass lies near q, a moderate temperature closely approximates the hard coverage event; when mass accumulates at q, every smooth diagnostic must pay an explicit ambiguity term.

Second, represent a local target displacement as $P_T - P_S = \sum_j a_j \nu_j + R$. If the target-local failure gain along probe ν_j is upper-aligned with source vulnerability v_j, and its amplitude a_j is dominated by the unlabeled proxy d_j, then

$$M(P_T; q) - M(P_S; q) \;\lesssim\; c\, R_{\mathrm{SCBS}} + \Delta_{\mathrm{align}} + \Delta_{\mathrm{proxy}} + \Delta_R + \Delta_{\mathrm{boundary}}.$$

The four remainder terms respectively expose mismatch between source probes and target mechanisms, error in measuring drift amplitude, unrepresented shift, and the soft-to-hard approximation. Thus a large SCBS value is evidence under alignment, while a small value is reassuring only to the extent that these terms are controlled.

Finally, source-only routing is a finite model-selection problem. If the two channels' population pseudo-deployment AUROCs differ by $\delta > 0$, the probability of choosing the inferior channel decreases exponentially in the number of independent benign–harmful comparison pairs (up to the dependence correction stated in Appendix A.7). This result justifies freezing a channel before observing the natural target stream; it does not assert that the simulation ranking must match every deployment. Appendix A states all assumptions and proofs. No claim of label-free validity under arbitrary joint shift is made.

5 Related Work and Positioning

Conformal prediction under distribution shift. Split conformal prediction turns held-out residuals or nonconformity scores into finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). Modern variants improve the adaptivity of classification sets (Romano et al., 2020), make image-classifier uncertainty sets practical (Angelopoulos et al., 2021), or control more general risks (Angelopoulos et al., 2024). Because ordinary exchangeability rarely survives deployment, another line of work modifies the conformal construction itself: importance weighting handles covariate shift when density ratios are estimable (Tibshirani et al., 2019), adaptive updates target time-varying sequences (Gibbs & Candès, 2021), and weighted or localized analyses relax exchangeability under additional structure (Barber et al., 2023). Unlabeled target data can also be used to adapt the set construction under additional modeling assumptions (Kasa et al., 2025). Recent results further clarify that conformal validity can be stated for arbitrary data distributions with an appropriately defined random test point (Prinster et al., 2024), while multi-source and cellwise-contamination settings require assumptions tailored to their shift mechanisms (Liu et al., 2024; Peng et al., 2025). These methods ask how to construct a valid or adaptive prediction set. We ask a complementary post-deployment question: given a *fixed* predictor, calibration set, and threshold, which unlabeled windows warrant concern that realized coverage has silently deteriorated while point accuracy remains stable? We neither change the set nor claim distribution-free validity under arbitrary changes in $P(Y \mid X)$.

Two-sample testing and shift detection. Kernel two-sample tests (Gretton et al., 2012) and classifier two-sample tests (Lopez-Paz & Oquab, 2017) can detect broad discrepancies between source and target covariates. Practical pipelines use dimension reduction followed by hypothesis tests to make distribution shift "fail loudly" (Rabanser et al., 2019). These are omnibus procedures: a benign seasonal change and a coverage-destroying change may both be easy to detect. Our target is therefore not the null $P_X = Q_X$, but the operational event that coverage falls by at least a chosen tolerance while point accuracy remains nearly unchanged. SCBS reweights observable drift by source-side evidence that the same feature direction moves probability mass across the conformal boundary. The C2ST and MMD comparisons in Section 6 empirically test whether this distinction matters after a persistent domain change has already occurred.

Unlabeled performance estimation and deployment monitoring. Predictive uncertainty often degrades under dataset shift even when standard in-domain metrics appear reliable (Ovadia et al., 2019). Several recent methods estimate whether a deployed classifier remains suitable without requiring a fully labeled target set. Mandoline estimates target performance from user-specified shift slices and limited structural assumptions (Chen et al., 2021); the Suitability Filter tests deployment requirements from unlabeled data plus source-side information (Pouget et al., 2025); retraining policies formalize when monitored performance justifies model replacement (Regol et al., 2025); and SHIFT hierarchically diagnoses which subgroups experience performance decay and attributes it to variable-specific shifts (Singh et al., 2025). WATCH uses weighted conformal martingales for sequential changepoint monitoring and diagnosis (Prinster et al., 2025). Our estimand and information setting are different: we rank unlabeled batches by coverage loss of a fixed conformal set, focus on failures hidden from ordinary accuracy, and localize evidence at the calibrated set boundary. The

optional uniform-label probe is reported separately because labels change the identification problem rather than merely improving an unlabeled score.

Feature attribution as a diagnostic primitive. Permutation importance and SHAP (Lundberg & Lee, 2017) describe how features affect prediction or average loss, but neither is designed around a conformal quantile. Feature-shift benchmarks such as TabFSBench (Cheng et al., 2025) likewise expose a broad range of marginal and dependency changes without declaring which changes should damage conformal coverage. SCBS uses the same feature-local vocabulary for a different purpose: an intervention is important only to the extent that it moves source examples across the fixed conformal boundary, and target drift matters only when it aligns with that vulnerability. This product structure is the main inductive bias evaluated in our controlled and natural-shift experiments.

6 EXPERIMENTS

The evaluation follows four questions implied by the proposed principle: whether boundary locality supplies signal beyond drift magnitude and predictive importance (RQ1), whether that signal transfers beyond the probes used to construct it (RQ2), whether it isolates harmful windows when generic shift detection saturates (RQ3), and what changes when target labels become available (RQ4). Appendix B gives the complete per-stream tables, ablations, negative controls, and implementation details.

6.1 SETUP

Controlled feature shifts. We use nine public numerical classification datasets, five random splits, and logistic regression, a two-layer MLP, and a 300-tree random forest. The 50/25/25 train/calibration/test split is fixed before applying missingness, permutation, noise, or rescaling to target features. Feature subsets cover high-, low-, and randomly selected importance groups. The Cartesian design has 57,600 scenarios. Unless stated otherwise, $\alpha = 0.10$ and a hidden failure means coverage gap $G \geq 0.05$ with accuracy drop at most 0.01.

Natural and external shifts. The frozen external grid uses five TabFSBench tasks, three models, three seeds, three nominal levels, and target batches of 128, 512, or the full test set (34,020 scenarios). Natural streams follow fixed source-to-target protocols: date-ordered Electricity, supplied-order OpenML Airlines, and date-ordered UCI Bank Marketing. All use non-overlapping target windows and three predictor seeds; the post-call duration field is removed from Bank. TableShift/NHANES, BRFSS, and Folktables ACS audits supply accuracy-visible and benign-drift controls.

Baselines and units. Controlled baselines are mean-score sensitivity, unweighted drift, permutation-importance times drift, localized SHAP concentration, and Output-UPE. For natural streams we add C2ST-HGB, a held-out nonlinear classifier two-sample test, and unbiased linear-time RBF-MMD. Each strong baseline uses a source-only 95th-percentile threshold from 60 source-vs-source resamples; C2ST uses at most 2,048 points per domain, while MMD uses a source median bandwidth and five pairings. Metrics are computed within dataset/model/window cells and then macro-averaged; correlated windows are never treated as independent replicates.

6.2 RQ1: IS BOUNDARY LOCALITY THE MISSING SIGNAL?

Among 25,183 accuracy-stable controlled scenarios, 981 (3.90%) are hidden five-point failures. SCBS obtains dataset-macro AUROC 0.846 and AUPRC 0.293 (Table 1); its paired dataset-bootstrap AUROC gains over Mean-CFSR, drift-only, and PFI$\times$Drift are respectively 0.102 (95% CI 0.038–0.181), 0.091 (0.042–0.150), and 0.220 (0.098–0.346). At the rarer ten-point endpoint, SCBS AUROC is 0.939.

Neither raw drift magnitude nor ordinary predictive importance is sufficient: drift-only loses 0.091 AUROC and PFI$\times$Drift loses 0.220 relative to SCBS. Mean-score sensitivity is closer but remains weaker, supporting localization at the calibrated boundary rather than sensitivity of the average score.

Table 1: Dataset-macro detection of controlled hidden coverage failures.

Method	AUROC	AUPRC	$\rho(G)$	AUROC, $G \geq .10$
SCBS	**0.846**	**0.293**	**0.340**	**0.939**
Mean-CFSR	0.745	0.207	0.248	0.835
Drift-only	0.756	0.137	0.235	0.818
PFI×Drift	0.626	0.080	0.143	0.659

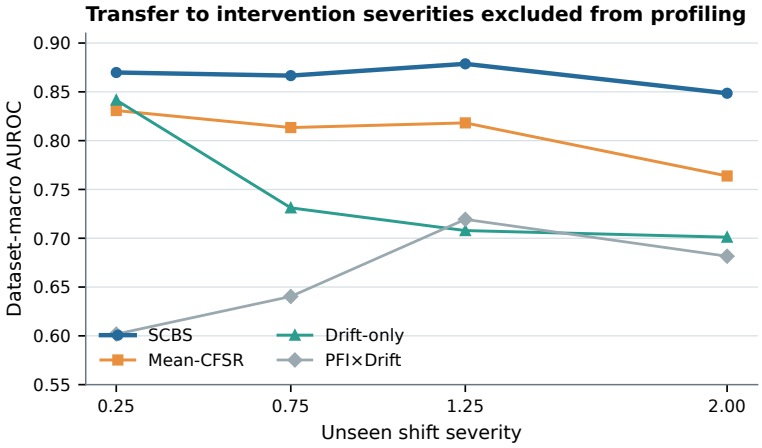


Figure 2: Transfer to unseen perturbation severities. Dataset-macro AUROC with paired dataset-level bootstrap intervals.

6.3 RQ2: Does alignment transfer beyond its probes?

Transfer tests preserve the ordering without fitting SCBS to held-out target labels. Leave-one-dataset-out AUROC is 0.833 versus 0.807, 0.804, and 0.722 for Mean-CFSR, drift-only, and PFI×Drift. Leaving out each shift family gives SCBS mean/worst-family AUROC 0.822/0.802. At unseen severity multipliers 0.25, 0.75, 1.25, and 2.0, AUROC is 0.870, 0.867, 0.879, and 0.849. Temperature, calibration-size, SHAP, correction, and runtime ablations are in Appendix B; notably, correction does not outperform alternative sensitivity features, so we claim detection rather than coverage repair.

The modest decline at multiplier 2.0 is consistent with severe shifts becoming accuracy-visible and boundary locality becoming less informative far from source support. The score is never refit at these held-out magnitudes.

External tasks. On the five-task TabFSBench grid at $\alpha = 0.10$, SCBS ranks first for every batch size, with AUROC 0.824, 0.824, and 0.811 for batches 128, 512, and full test. At $\alpha = 0.05$, PFI×Drift is stronger, an explicit non-uniformity. Folktables ACS contains no accuracy-stable five-point failures and Dir-SimCal raises no alarms; TableShift/NHANES contains accuracy-visible failures for which the output channel is stronger. These retained negative regimes prevent the natural hidden endpoint from being confused with generic performance degradation.

6.4 RQ3: Can monitoring separate harmful from detectable shift?

Electricity and Airlines contain complementary hidden-failure mechanisms: Dir-SimCal is strongest on Electricity (AUROC/AUPRC 0.969/0.901), whereas Output-UPE is strongest on Airlines (0.988 ± 0.002/0.828 ± 0.041). We fixed a router using only labeled source pseudo-deployments from those two streams: within each model/level/window cell, it selects the normalized channel with larger harmful-versus-benign source AUROC, falling back to Output-UPE when fewer than five harmful simulations exist. Bank Marketing was then evaluated score-blind, without changing the rule. The router obtains Bank AUROC/AUPRC 0.974 ± 0.037/0.997 ± 0.003, recall 0.828, FPR 0.159, and precision 0.951.

Table 2: Three natural hidden-failure streams. Values are dataset-macro mean ± standard deviation over three predictor seeds; alarms use only source-selected thresholds.

Score	AUROC	AUPRC	Recall	FPR	Precision
C2ST-HGB	0.582 ± 0.018	0.343 ± 0.030	1.000	1.000	0.242
Linear MMD	0.746 ± 0.029	0.529 ± 0.002	1.000	0.980	0.244
Generic drift	0.790 ± 0.024	0.566 ± 0.004	1.000	1.000	0.242
Dir-SimCal	0.883 ± 0.026	0.811 ± 0.009	0.652	**0.120**	0.535
Output-UPE	0.808 ± 0.028	0.688 ± 0.031	0.770	0.173	0.481
Dual-path fusion	0.965 ± 0.018	0.825 ± 0.025	**0.864**	0.203	0.503
Source router	$\mathbf{0.965 \pm 0.025}$	$\mathbf{0.894 \pm 0.035}$	0.840	0.146	**0.640**

Table 2 makes the distinction from omnibus shift testing: C2ST-HGB detects every domain change but has AUROC 0.582 for harmful windows, while MMD flags 98.0% of benign windows. The router preserves fusion's AUROC, raises AUPRC to 0.894, and lowers FPR to 0.146; Bank moving-block intervals confirm this improvement (Appendix B).

6.5 RQ4: What do labels add?

A uniform target-label probe changes the information setting. Across 30 nested draws per Electricity window, the model-assisted 128-label estimator obtains AUROC/AUPRC 0.995/0.956. Screening recall/FPR are 0.767/0.017, while strictly certifying $G > 0.05$ yields 0.272/0.0005. Labels therefore enable a strong screen, not high-power certification. Target profiling takes 13.42 ms per batch and aggregation 1.55 μs. Bank is score-blind, but the three screened streams support a warning rather than universal validity.

7 Limitations

SCBS and Dir-SimCal are warnings, not coverage certificates. They cannot identify label mechanisms invisible in X, interactions outside the probes, or small-window drift. These correspond to the theory's probe-mismatch, drift-proxy, unrepresented-shift, and boundary-approximation terms. Three natural streams do not establish universal transfer: the channels are regime dependent, pseudo-shifts need not span a new domain, Airlines is not timestamped, and Bank cells contain only 4–12 stable windows. Label probes consume up to half the smallest window and remain low-power for certification; source profiling requires repeated offline evaluations.

8 Conclusion

A conformal predictor can lose coverage while its classifier remains accurate, yet generic shift detection does not reveal whether the change is harmful. Because coverage is not identifiable from unlabeled covariates under arbitrary joint shift, we introduce boundary alignment as a structured inductive bias: SCBS combines where deployment moved with source evidence about which directions cross the fixed conformal threshold. Controlled, external, and natural evaluations support the principle while exposing its limits: C2ST and MMD detect broad change but not harmful windows selectively. Calibration and routing improve operation without strengthening validity; target labels instead close the identification loop. Thus the framework separates certification, structured warning, and label-assisted identification.

AI use. Generative AI helped retrieve candidate related work and organize search terms; every citation was manually verified, and the authors take responsibility for the paper.

Reproducibility. An anonymized code archive is supplied as supplementary material and at `https://anonymous.4open.science/r/SCBS-E81F`; it contains implementations, tests, environment files, entry points, and public-data instructions, with full protocol details in Section 6 and Appendices B–C.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

F. Jay Breidt and Jean D. Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205, 2017. doi: 10.1214/16-STS589.

Mayee F. Chen, Karan Goel, Nimit S. Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1617–1629. PMLR, 2021.

Zi-Jian Cheng, Ziyi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. TabFSBench: Tabular benchmark for feature shifts in open environments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10025–10089, 2025.

Frances Ding, Moritz Hardt, John P. Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6478–6490, 2021.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2324.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

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

Kevin Kasa, Zhiyu Zhang, Heng Yang, and Graham W. Taylor. Adapting prediction sets to distribution shifts without labels. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1990–2010. PMLR, 2025.

Yi Liu, Alexander Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31344–31382. PMLR, 2024.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing. UCI Machine Learning Repository, 2014.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Qian Peng, Yajie Bao, Haojie Ren, Zhaojun Wang, and Changliang Zou. Conformal prediction with cellwise outliers: A detect-then-impute approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48808–48831. PMLR, 2025.

Angéline Pouget, Mohammad Yaghini, Stephan Rabanser, and Nicolas Papernot. Suitability filter: A statistical framework for classifier evaluation in real-world deployment settings. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49603–49627. PMLR, 2025.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41086–41118. PMLR, 2024.

Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49830–49859. PMLR, 2025.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Florence Regol, Leo Schwinn, Kyle Sprague, Mark Coates, and Thomas Markovich. When to retrain a machine learning model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51369–51404. PMLR, 2025.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Harvineet Singh, Fan Xia, Alexej Gossmann, Andrew Chuang, Julian C. Hong, and Jean Feng. "Who experiences large model decay and why?" a hierarchical framework for diagnosing heterogeneous performance drift. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55757–55787. PMLR, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

Alexander Vergara. Gas sensor array drift dataset. UCI Machine Learning Repository, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

A THEORETICAL INTERPRETATION OF SCBS

A.1 SETUP

Let P_S and P_T denote the source and target distributions over (X, Y). Let a fixed probabilistic predictor induce the nonconformity score

$$S = s(X, Y) = 1 - p_\theta(Y \mid X),$$

and let q be a fixed split-conformal threshold. Throughout the theoretical statements below, q is treated as fixed; equivalently, one may condition on an independently estimated threshold or use a cross-fitted analysis split.

Define the hard miscoverage functional

$$M(P; q) = \Pr_{(X,Y)\sim P}\{S > q\},$$

so that

$$C(P; q) = 1 - M(P; q).$$

For $\tau > 0$, define the soft boundary indicator

$$\phi_\tau(s; q) = \sigma\left(\frac{s - q}{\tau}\right)$$

and soft miscoverage

$$M_\tau(P; q) = \mathbb{E}_P\left[\phi_\tau(S; q)\right].$$

A.2 AN UNLABELED IMPOSSIBILITY RESULT

The following construction makes the information limitation explicit and is independent of how a warning score is parameterized.

Proposition 2 (Unlabeled covariates do not identify target coverage)**.** *Fix a finite label space $\mathcal{Y}$, a prediction-set rule $\widehat{C} : \mathcal{X} \to 2^{\mathcal{Y}}$, and a target marginal Q_X. Suppose there is a measurable set A with $Q_X(A) > 0$ such that $\emptyset \neq \widehat{C}(x) \neq \mathcal{Y}$ for every $x \in A$, and suppose measurable selectors $y_{\mathrm{in}}(x) \in \widehat{C}(x)$ and $y_{\mathrm{out}}(x) \notin \widehat{C}(x)$ exist on A. Then there are target joint distributions Q_0 and Q_1 with the same covariate marginal Q_X whose coverages differ by $Q_X(A)$. Moreover, every statistic measurable with respect to source data and unlabeled target covariates has the same distribution under Q_0 and Q_1.*

Proof. Choose any common conditional label distribution outside A. On A, let $Q_0(Y = y_{\mathrm{in}}(X) \mid X) = 1$ and $Q_1(Y = y_{\mathrm{out}}(X) \mid X) = 1$. Both joint distributions have marginal Q_X. Their coverage indicators agree outside A, equal one under Q_0 on A, and equal zero under Q_1 on A; hence their coverages differ by $Q_X(A)$. Since source observations are fixed by the experimental setup and the target covariates have the same law, the observable sigma-field has the same law in both worlds. Therefore no label-free statistic can distinguish them. □

The proposition rules out a universal label-free certificate, not useful warning under structure. SCBS introduces such structure through the probe-alignment and drift-dominance assumptions below; the optional target label probe changes the information set.

A.3 SOFT MISCOVERAGE APPROXIMATES HARD MISCOVERAGE

Lemma 3 (Boundary-mass approximation)**.** *For every $\epsilon > 0$,*

$$|M_\tau(P; q) - M(P; q)| \leq \Pr_P\{|S - q| \leq \epsilon\} + e^{-\epsilon/\tau}.$$

Proof. If $S \geq q + \epsilon$, then

$$1 - \phi_\tau(S; q) = \frac{1}{1 + \exp((S - q)/\tau)} \leq e^{-\epsilon/\tau}.$$

If $S \leq q - \epsilon$, then

$$\phi_\tau(S; q) \leq e^{-\epsilon/\tau}.$$

Inside the band $|S - q| \leq \epsilon$, the pointwise discrepancy from the hard indicator is at most one. Taking expectations yields the result. □

Corollary 4 (Density-controlled soft-to-hard error)**.** *Assume the score distribution under P admits a density bounded by L on $[q - \epsilon, q + \epsilon]$. Then*

$$|M_\tau(P; q) - M(P; q)| \leq 2L\epsilon + e^{-\epsilon/\tau}.$$

Hence, whenever the score distribution has bounded local density around q,

$$M_\tau(P; q) \to M(P; q) \qquad \text{as } \tau \to 0.$$

A.4 What the source-side SCBS probe measures

Let $P_S^{\pi_j}$ denote the source distribution after replacing feature X_j by an independent draw from its source marginal while keeping (X_{-j}, Y) unchanged. This is the population analogue of the feature permutation used by the algorithm.

Define the hard intervention-induced increase in miscoverage

$$\Delta_j = M(P_S^{\pi_j}; q) - M(P_S; q)$$

and the corresponding soft quantity

$$\Delta_j^\tau = M_\tau(P_S^{\pi_j}; q) - M_\tau(P_S; q).$$

The population SCBS vulnerability is

$$v_j = [\Delta_j^\tau]_+.$$

Proposition 5 (SCBS approximates hard intervention vulnerability)**.** *Define*

$$B_\tau(P; q, \epsilon) = \Pr_P\{|S - q| \leq \epsilon\} + e^{-\epsilon/\tau}.$$

Then

$$\left|[\Delta_j^\tau]_+ - [\Delta_j]_+\right| \leq B_\tau(P_S^{\pi_j}; q, \epsilon) + B_\tau(P_S; q, \epsilon).$$

Thus, when both the original and intervened score laws place little mass near the conformal boundary, SCBS closely approximates the true increase in miscoverage caused by breaking feature j's source relation.

Proof. The map $x \mapsto [x]_+$ is 1-Lipschitz, so

$$\left|[\Delta_j^\tau]_+ - [\Delta_j]_+\right| \leq |\Delta_j^\tau - \Delta_j|.$$

Apply the boundary-mass lemma separately to $P_S^{\pi_j}$ and P_S, then use the triangle inequality. □

A.5 A target-shift decomposition that does not assume permutation shifts

The target distribution need not be obtainable by literal source feature permutations. Instead, model the target displacement in the space of signed measures as

$$P_T - P_S = \sum_{j=1}^{d} a_j \nu_j + R, \qquad a_j \geq 0,$$

where:

- ν_j is a normalized signed measure representing a feature-local target perturbation direction associated with feature j;

- a_j is the latent magnitude of target shift along that direction;
- R is a residual signed measure containing interactions, joint shifts, and other components not represented by the feature-local dictionary.

This decomposition is a local/tangent representation of the target shift, not an assumption that the target distribution is literally a mixture of permuted source distributions.

Define the target-direction soft failure gain

$$g_j = \int \phi_\tau(s(x,y);q)\, d\nu_j(x,y).$$

We use two explicit assumptions to connect the deployable statistic to the latent target shift.

Assumption A1: probe alignment. There exists $\kappa_v > 0$ and feature-wise alignment slack $\xi_j \geq 0$ such that

$$g_j \leq \kappa_v v_j + \xi_j \qquad \text{for all } j.$$

This states that the source permutation probe is sufficiently sensitive to target-local directions that can increase conformal miscoverage, while ξ_j explicitly records probe mismatch. The assumption does not require the target perturbation itself to be a permutation.

Assumption A2: unlabeled drift dominance. Let $d_j \geq 0$ denote the source-null-calibrated unlabeled target drift score used by the algorithm. There exist $\kappa_d > 0$ and $\eta_j \geq 0$ such that

$$a_j \leq \kappa_d d_j + \eta_j.$$

The common multiplicative constant permits the drift statistic to have an arbitrary global scale; η_j captures feature-wise proxy error.

Theorem 6 (SCBS risk is an upper surrogate for soft coverage loss)**.** *Under Assumptions A1–A2,*

$$\begin{aligned} M_\tau(P_T;q) - M_\tau(P_S;q) \leq \kappa_v\kappa_d \sum_{j=1}^{d} v_j d_j \\ + \kappa_v \sum_{j=1}^{d} v_j \eta_j \\ + \sum_{j=1}^{d} a_j \xi_j \\ + \|R\|_{\mathrm{TV}}. \end{aligned}$$

Therefore, defining

$$R_{\mathrm{SCBS}} = \sum_{j=1}^{d} v_j d_j,$$

the proposed risk controls the represented component of target soft miscoverage increase up to a global scale, drift-proxy error, and unrepresented residual shift.

Proof. Because $M_\tau(P;q)$ is linear in P for fixed predictor and fixed q,

$$M_\tau(P_T;q) - M_\tau(P_S;q) = \sum_{j=1}^{d} a_j g_j + \int \phi_\tau\, dR.$$

Since $\|\phi_\tau\|_\infty \leq 1$,

$$\left| \int \phi_\tau\, dR \right| \leq \|R\|_{\mathrm{TV}}.$$

Using $a_j \geq 0$ and Assumption A1,

$$\sum_j a_j g_j \leq \kappa_v \sum_j a_j v_j + \sum_j a_j \xi_j.$$

Applying Assumption A2 gives

$$\sum_j a_j v_j \le \kappa_d \sum_j v_j d_j + \sum_j v_j \eta_j.$$

Combining the inequalities yields the result. □

Corollary 7 (Hard coverage-loss interpretation)**.** *For any $\epsilon > 0$,*

$$\begin{aligned} C(P_S; q) - C(P_T; q) \le{}& \kappa_v \kappa_d R_{\mathrm{SCBS}} \\ &+ \kappa_v \sum_j v_j \eta_j \\ &+ \sum_j a_j \xi_j \\ &+ \|R\|_{\mathrm{TV}} \\ &+ E_{\mathrm{bdry}}, \end{aligned}$$

where

$$E_{\mathrm{bdry}} = B_\tau(P_S; q, \epsilon) + B_\tau(P_T; q, \epsilon).$$

Thus the hard target coverage loss is bounded by the SCBS risk plus four explicit approximation terms:

$$\begin{aligned} &\textit{drift-proxy error} + \textit{probe-alignment error} \\ &\quad + \textit{unrepresented shift} + \textit{soft-boundary error}. \end{aligned}$$

A.6 Finite-sample estimation: a cross-fitted theoretical variant

The main algorithm estimates v_j by repeated permutations of a finite source calibration set. To state a clean finite-sample result without hiding dependencies introduced by reusing the same calibration sample for both q and vulnerability estimation, consider the following cross-fitted theoretical variant:

1. estimate q on an independent calibration fold;
2. estimate v_j on a separate vulnerability fold;
3. for each observation, replace X_j by an independent draw from the population source marginal or from an independent source reference pool.

Let

$$Z_{ij} = \phi_\tau(S_{ij}^\pi; q) - \phi_\tau(S_i; q).$$

Then $Z_{ij} \in [-1, 1]$. Define

$$\widehat{\Delta}_j^\tau = \frac{1}{n} \sum_{i=1}^n Z_{ij}, \qquad \widehat{v}_j = [\widehat{\Delta}_j^\tau]_+.$$

Proposition 8 (Uniform concentration for cross-fitted SCBS)**.** *For the independent-fold estimator above, with probability at least $1 - \delta$,*

$$\max_{1 \le j \le d} |\widehat{v}_j - v_j| \le \sqrt{\frac{2 \log(2d/\delta)}{n}}.$$

Proof. For each feature, Hoeffding's inequality for variables in $[-1, 1]$ gives

$$\Pr\left(\left|\widehat{\Delta}_j^\tau - \Delta_j^\tau\right| > t\right) \le 2e^{-nt^2/2}.$$

A union bound over d features and the 1-Lipschitz property of $x \mapsto [x]_+$ yield the result. □

Relation to the implemented estimator. The implementation reuses the calibration sample and averages repeated column permutations for efficiency. The concentration proposition is therefore presented as a clean theoretical variant rather than claimed verbatim for the implemented finite-sample estimator. Empirically, the calibration-size sensitivity study evaluates the stability of the implemented estimator directly.

A.7 FINITE-SAMPLE SOURCE-ROUTER SELECTION

Consider two fixed score channels evaluated on independent source pseudo-deployments. Let A_1 and A_2 be their population harmful-versus-benign AUROCs, with $A_1 - A_2 = \Delta > 0$. Let $\widehat{A}_k$ be the empirical AUROC computed from n_+ iid harmful and n_- iid benign pseudo-deployments, and select the channel with larger $\widehat{A}_k$.

Proposition 9 (Probability of selecting the inferior channel)**.** *Conditional on the fitted scores and the iid pseudo-deployment generator,*

$$\Pr(\widehat{A}_2 \geq \widehat{A}_1) \leq 4 \exp\left\{ -\frac{\Delta^2}{2(1/n_+ + 1/n_-)} \right\}.$$

Proof. Empirical AUROC is an average of $n_+ n_-$ pairwise ranking indicators. Replacing one harmful observation changes it by at most $1/n_+$, and replacing one benign observation changes it by at most $1/n_-$. McDiarmid's inequality therefore gives $\Pr(|\widehat{A}_k - A_k| \geq t) \leq 2\exp\{-2t^2/(1/n_+ + 1/n_-)\}$. Wrong selection implies that at least one channel deviates from its population AUROC by $\Delta/2$ in the adverse direction. A union bound gives the claim. □

This result explains how a genuine source-simulation AUROC margin stabilizes routing as the pseudo-deployment bank grows. It is conditional on the simulation distribution and does not guarantee that the source ordering transfers to an arbitrary natural target stream.

A.8 INTERPRETATION AND SCOPE

The theorem explains the multiplicative structure

$$R_{\mathrm{SCBS}} = \sum_j \underbrace{v_j}_{\text{source conformal vulnerability}} \underbrace{d_j}_{\text{unlabeled target drift}}.$$

Large feature drift alone is insufficient: a feature contributes strongly only if deployment drift occurs along a direction to which the conformal boundary is vulnerable.

The result is a diagnostic-surrogate statement, not an unconditional distribution-free target-coverage guarantee. Under arbitrary joint shift of (X, Y), unlabeled target covariates do not in general identify the target score distribution. The role of Assumptions A1–A2 is to make explicit the structural conditions under which unlabeled drift can inform conformal failure risk.

B EXTENDED EXPERIMENTAL RESULTS

We organize the evaluation around eight questions. **RQ1:** Can SCBS detect material conformal coverage loss, especially when accuracy appears stable? **RQ2:** Does boundary sensitivity add information beyond generic drift and predictive feature importance? **RQ3:** Does the diagnostic transfer across datasets, models, shift families, and unseen severities? **RQ4:** Is the result robust to the temperature and calibration-sample size? **RQ5:** Does it transfer to independent benchmarks and natural domain shifts? **RQ6:** How does direction-aware scoring compare and combine with unlabeled output-based estimation? **RQ7:** What is gained when a small target-label budget is available? **RQ8:** Is the computation practical at deployment?

B.1 EXPERIMENTAL SETUP

Data and predictors. We use nine public numerical classification datasets: Breast Cancer, Wine, and Digits from scikit-learn, and Spambase, Ionosphere, Sonar, Diabetes, Vehicle, and Banknote Authentication from OpenML (Pedregosa et al., 2011; Vanschoren et al., 2014). They span binary and multiclass tasks and 4–64 features. Each dataset is stratified into 50% training, 25% calibration, and 25% test partitions. We repeat the complete split with five seeds. Predictors are logistic regression, a two-hidden-layer MLP, and a 300-tree random forest, covering linear, neural, and tree-based decision rules.

Table 3: Dataset-macro detection of hidden conformal failures (accuracy drop ≤ 0.01). The number of valid datasets is smaller for the ten-point endpoint when a dataset has only one class of outcomes.

Failure	Method	Datasets	AUROC	AUPRC	Spearman
$G \geq 0.05$	SCBS	9	**0.846**	**0.293**	**0.340**
	Mean-CFSR	9	0.745	0.207	0.248
	Drift-only	9	0.756	0.137	0.235
	PFI×Drift	9	0.626	0.080	0.143
$G \geq 0.10$	SCBS	7	**0.939**	**0.245**	**0.340**
	Mean-CFSR	7	0.835	0.137	0.248
	Drift-only	7	0.818	0.060	0.235
	PFI×Drift	7	0.659	0.031	0.143

Feature shifts and scenarios. We intervene on test features using four families: mean-imputed missingness, independent permutation, additive noise, and multiplicative rescaling. For each dataset/model/seed cell, feature subsets include the most and least important subsets and 30 random subsets at feasible subset sizes. The source model and conformal threshold are never refit after a shift. The Cartesian design contains 57,600 target scenarios. Our principal analysis fixes $\alpha = 0.10$ and uses the clean split-conformal predictor as the object being monitored.

Diagnostics. We compare SCBS with (i) **Mean-CFSR**, which replaces the soft boundary change with mean nonconformity-score change; (ii) **Drift-only**, the unweighted sum of target drift; (iii) **PFI×Drift**, permutation importance multiplied by the same drift profile; (iv) localized SHAP concentration and SHAP-concentration×drift; and (v) **Output-UPE**, a suitability-signal estimator trained on labeled source monitoring data to predict conformal misses from confidence, margin, entropy, set size, and set mass. Its target score is the increase in predicted miss rate and its alarm threshold is source-bootstrapped. Accuracy drop is reported only as a labeled retrospective oracle. Unless noted otherwise, SCBS uses the fixed temperature scale τ obtained from the 20th percentile of $|S - q|$ on the source calibration fold.

On the three natural hidden-failure streams we additionally compare two omnibus covariate-shift detectors. **C2ST-HGB** is a balanced, held-out classifier two-sample test using histogram gradient boosting; **Linear MMD** is the unbiased linear-time RBF-MMD estimator with a source median-heuristic bandwidth. Their alarm thresholds are source-only 95th percentiles from 60 source-vs-source resamples. The C2ST uses at most 2,048 observations per domain and MMD averages five random pairings.

Metrics and statistical unit. The primary endpoint is dataset-macro AUROC for detecting $G \geq 0.05$ among scenarios whose accuracy drop is at most 0.01. We also report AUPRC, Spearman correlation with the continuous gap, and detection of $G \geq 0.10$. We compute every metric within a dataset before macro-averaging, so the 57,600 correlated scenarios are not treated as independent replicates. Uncertainty for method differences is obtained by a paired bootstrap over datasets (10,000 resamples). Leave-one-dataset-out (LODO), leave-one-shift-family-out (LFO), and leave-one-severity-out analyses test transfer rather than in-sample fit.

B.2 Main result: hidden coverage failures

Table 3 reports the primary hidden-failure result. SCBS reaches macro AUROC 0.846 for five-point failures and 0.939 for ten-point failures. At the five-point threshold, it improves over Mean-CFSR, drift-only, and PFI×Drift by 0.102, 0.091, and 0.220 AUROC, respectively. The AUPRC values are lower because hidden severe failures are rare: among the 25,183 scenarios satisfying the accuracy constraint, 3.90% have $G \geq 0.05$ and 0.73% have $G \geq 0.10$. For this reason, both ranking metrics are needed.

The paired dataset bootstrap supports that these are not driven by a single benchmark. For $G \geq 0.05$, the AUROC differences versus Mean-CFSR, drift-only, and PFI×Drift are 0.102 (95% CI 0.038–0.181), 0.091 (0.042–0.150), and 0.220 (0.098–0.346). The corresponding AUPRC intervals also

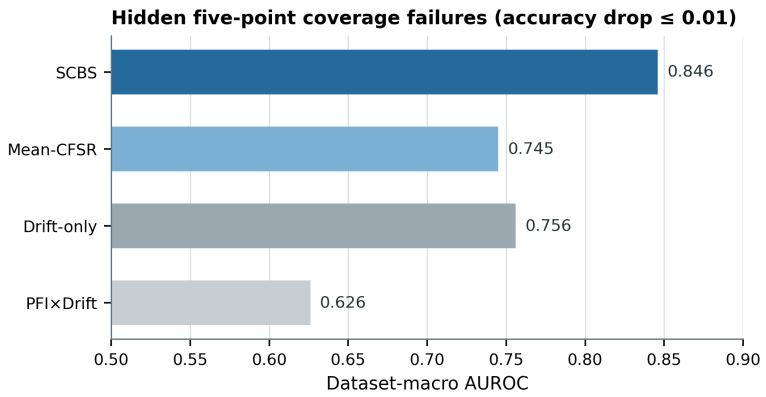


Figure 3: Dataset-macro AUROC for detecting a coverage gap of at least five points among scenarios with accuracy drop at most one point.

Table 4: Unseen-severity generalization for hidden five-point failures.

Method	0.25	0.75	1.25	2.00
SCBS	**0.870**	**0.867**	**0.879**	**0.849**
Mean-CFSR	0.831	0.813	0.818	0.764
Drift-only	0.842	0.731	0.708	0.701
PFI×Drift	0.602	0.640	0.719	0.682

exclude zero. For $G \geq 0.10$, all six AUROC/AUPRC comparisons exclude zero; the AUROC gain over PFI×Drift is 0.281 (0.175–0.400). Figure 3 visualizes the main endpoint.

B.3 Transfer across datasets and shift families

LODO auditing preserves the ordering of methods. A fold-wise audit gives macro AUROC 0.833 for SCBS versus 0.807 for Mean-CFSR, 0.804 for drift-only, and 0.722 for PFI×Drift on the hidden five-point endpoint. The analysis fits only the orientation/calibration of a scalar score on the remaining datasets; the raw SCBS construction itself has no target labels.

In leave-one-shift-family-out evaluation, the fixed source-side permutation probe is evaluated on a corruption family excluded from score selection. SCBS obtains mean-shift LFO AUROC 0.822 for hidden five-point failures, with worst-family AUROC 0.802. The source-selected narrower boundary variant reaches 0.850 mean and 0.818 worst-family AUROC. These results support, but do not prove, the probe-alignment assumption: a source permutation probe transfers to missingness, noise, rescaling, and held-out permutation scenarios.

B.4 Generalization to unseen severity

We compute the source profile at the reference severity and evaluate previously unseen severities 0.25, 0.75, 1.25, and 2.0. For hidden five-point failures, SCBS macro AUROC is 0.870, 0.867, 0.879, and 0.849, respectively (Table 4). It is the best method at all four severities; the smallest margin is at severity 0.25, where drift-only is competitive. Performance does not increase mechanically with corruption magnitude, which is expected because the hidden-failure filter selects shifts with little accuracy degradation.

B.5 Boundary and sample-size ablations

SCBS is robust to a broad temperature range, but a narrow boundary is consistently preferable for hidden failures. With temperature multipliers 0.25, 0.5, 1, 2, and 4, hidden-failure LODO AUROC is 0.821, 0.800, 0.775, 0.757, and 0.738. Mean-CFSR obtains 0.670. Because selecting the multiplier using target labels would leak evaluation information, all formal comparisons in Table 3 use the fixed multiplier 1; a separate source-only selection experiment is reported in Appendix C.

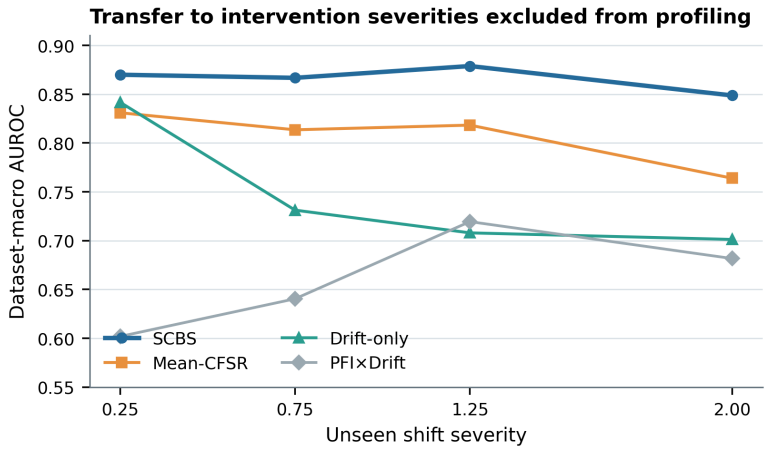


Figure 4: Macro AUROC across unseen shift severities for hidden five-point failures.

Calibration-size resampling uses 25%, 50%, 75%, and 100% of the available calibration fold with three repetitions. For the all-scenario five-point endpoint, the 0.25-temperature variant obtains LODO AUROC 0.861 ± 0.004, 0.848 ± 0.008, 0.847 ± 0.008, and 0.842 as the calibration fraction grows. The absence of monotone improvement indicates that the effect is not an artifact of a single large calibration set; the increasing zero-vulnerability rate at small sample sizes is documented in the appendix.

B.6 Alternative explanations and corrective methods

SCBS remains stronger than localized SHAP concentration: in the leakage-audited hidden five-point analysis, macro AUROC is 0.833 for SCBS, 0.548 for SHAP concentration, and 0.778 for SHAP-concentration×drift. Thus concentration of ordinary feature attribution does not explain the result.

We also test whether the diagnostic can drive threshold inflation. Isotonic mappings from risk to correction reduce average shortfall, but SCBS-based correction is not distinguishable from PFI- or mean-sensitivity-based correction: mean shortfall is 0.00911 for SCBS-Iso, 0.00912 for PFI-Iso, and 0.00908 for Mean-Iso, compared with 0.01033 for drift-only and 0.04713 for uncorrected SplitCP. We therefore treat adaptation as a secondary negative result and do not claim SCBS is a superior coverage-repair method. Estimated weighted conformal and KMM baselines are included in the appendix; both are unstable in the most severe feature-corruption regimes.

B.7 Independent benchmarks and natural shifts

We freeze the pipeline and evaluate five classification tasks from TabFSBench (Cheng et al., 2025): Heart, Electricity, Eye Movements, MiniBooNE, and Jannis. The external grid adds histogram gradient boosting to logistic regression and random forests, uses three seeds, varies $\alpha \in \{0.05, 0.10, 0.20\}$ and target batch size in $\{128, 512, \text{all}\}$, and contains 34,020 scenarios. Table 5 reports dataset-macro AUROC for detecting a five-point coverage loss without filtering on accuracy. At $\alpha = 0.10$, SCBS ranks first for every batch size. At $\alpha = 0.05$, however, PFI×Drift is stronger; the advantage of boundary sensitivity is therefore not uniform across nominal coverage levels.

Hidden failures are rare in the external full-test batches: for $\alpha = 0.10$, only one of 885 accuracy-stable scenarios has $G \geq 0.05$. Batch 128 yields 37 such failures among 963 accuracy-stable scenarios, but SCBS macro AUROC is only 0.631 and does not beat every comparator. We therefore use the external benchmark as evidence for overall ranking transfer, not as confirmation of the primary hidden-failure endpoint.

For a shift not constructed by our intervention generator, we order all 45,312 Electricity examples by their recorded date and period. The first 40% train the predictor, the next 10% calibrate conformal prediction and the drift reference, the next 10% define source-period performance, and the final 40% form non-overlapping windows of 256, 512, or 1,024 observations. Across three models and three α

Table 5: TabFSBench external validation: dataset-macro AUROC for $G \geq 0.05$.

α	Batch	SCBS	Mean-CFSR	PFI×Drift	Drift-only
0.05	128	0.772	0.749	**0.823**	0.617
	512	0.759	0.731	**0.816**	0.606
	all	0.749	0.717	**0.803**	0.613
0.10	128	**0.824**	0.802	0.809	0.651
	512	**0.824**	0.797	0.795	0.641
	all	**0.811**	0.778	0.777	0.652
0.20	128	**0.850**	0.829	0.801	0.673
	512	**0.867**	0.833	0.793	0.680
	all	**0.854**	0.811	0.778	0.689

Table 6: Electricity natural temporal audit at $\alpha = 0.10$ for hidden $G \geq 0.05$ failures. Ranking values are group-macro; alarm metrics use source-selected thresholds.

Method	AUROC	AUPRC	Recall	FPR	Precision
SCBS, source null	0.715	0.402	1.000	1.000	0.081
Directional, source null	**0.969**	**0.901**	1.000	1.000	0.081
Output-UPE	0.832	0.549	0.688	0.097	0.343
Dir-SimCal	**0.969**	**0.901**	0.833	**0.048**	**0.567**
Dual-path	0.942	0.624	**0.958**	0.110	0.444

values, 44 of 1,098 window-level evaluations exhibit $G \geq 0.05$ while accuracy drops by at most one point. This establishes that hidden coverage failures occur under natural temporal drift. The base 95th-percentile source-bootstrap alarm nevertheless flags every later window.

The direction-aware extension changes this result materially at the principal $\alpha = 0.10$ setting (Table 6). For hidden five-point failures, its group-macro AUROC/AUPRC is 0.969/0.901, compared with 0.715/0.402 for SCBS and 0.832/0.549 for Output-UPE. Source-simulation calibration retains 0.833 recall while reducing false-positive rate from 1.000 to 0.048 and raising precision from 0.081 to 0.567. When accuracy-visible windows are retained, recall is 0.812, false-positive rate 0.103, and precision 0.763. Each model/window cell uses 200 prespecified source pseudo-deployments; target labels do not select the score or threshold.

The dual-path score divides Dir-SimCal and Output-UPE by their respective source-null 97.5th percentiles and takes their maximum; Bonferroni correction gives a source-only 5% family-wise operating point. It is a high-sensitivity option rather than a uniform replacement: on hidden Electricity failures it raises recall from 0.833 to 0.958 but also raises FPR from 0.048 to 0.110 and weakens ranking.

We then seek a second natural hidden-failure stream without tuning the monitor on target labels. OpenML Airlines data set 1169 contains 539,383 flight records with the native binary delay target (Vanschoren et al., 2014). We preserve the supplied row order and apply the same 40/10/10/40 protocol, using non-overlapping windows of 512, 2,048, and 8,192 records. A prevalence-only screen fixes these choices before the monitoring analysis. Across three model seeds at $\alpha = 0.10$, the accuracy-stable subset contains 914–929 windows and 216–218 hidden five-point failures. Thus the endpoint is no longer confined to Electricity.

Table 7 reports mean and sample standard deviation over seeds; within each seed, metrics are first macro-averaged over model–window cells. Output-UPE ranks the Airlines failures almost perfectly. Dir-SimCal improves substantially over SCBS and reduces FPR from 1.000 to 0.136, but does not dominate the output channel. The dual path inherits Output-UPE's ranking and raises FPR. This is a genuine accuracy-stable counter-regime: hiddenness alone does not determine which unlabeled view is informative.

The complementary behavior on Electricity and Airlines motivated a source-only router, described fully in Appendix C.2. We fixed its normalization, harmful-source threshold, minimum harmful count, and fallback before selecting a third stream. UCI Bank Marketing provides 41,188 examples in date order from May 2008 to November 2010 (Moro et al., 2014). We remove the post-call duration

Table 7: OpenML Airlines supplied-order audit at $\alpha = 0.10$ for hidden $G \geq 0.05$ failures. Entries are mean $\pm$ standard deviation over three seeds; 6–7 model–window cells per seed contain both outcome classes for ranking metrics.

Method	AUROC	AUPRC	Recall	FPR	Precision
SCBS, source null	0.649 ± 0.012	0.359 ± 0.030	1.000	1.000	0.183
Dir-SimCal	0.818 ± 0.032	0.580 ± 0.053	0.642	**0.136**	**0.425**
Output-UPE	$\mathbf{0.988 \pm 0.002}$	$\mathbf{0.828 \pm 0.041}$	1.000	0.225	0.405
Dual-path	0.988 ± 0.002	0.828 ± 0.041	1.000	0.247	0.385

Table 8: Frozen UCI Bank Marketing confirmation at $\alpha = 0.10$ for hidden $G \geq 0.05$ failures. Entries are mean $\pm$ standard deviation over three seeds after model–window macro-averaging.

Method	AUROC	AUPRC	Recall	FPR	Precision
Dir-SimCal	0.895 ± 0.127	0.961 ± 0.063	0.464	0.177	0.590
Output-UPE	0.532 ± 0.147	0.615 ± 0.125	0.500	0.204	0.633
Dual-path fusion	0.962 ± 0.045	0.981 ± 0.015	0.639	0.250	0.661
Source router	$\mathbf{0.974 \pm 0.037}$	$\mathbf{0.997 \pm 0.003}$	**0.828**	**0.159**	**0.951**

variable, apply the same 40/10/10/40 split, and use non-overlapping windows of 256, 512, and 1,024 records. A frozen endpoint-only screen identified hidden failures without comparing monitor scores. Across seeds, the accuracy-stable subset contains 75 windows and 30–37 failures per seed across model–window cells.

Table 8 is the resulting frozen, score-blind third-stream evaluation. The router selects Dir-SimCal for every logistic-regression cell and Output-UPE for every random-forest and gradient-boosting cell. It reaches AUROC/AUPRC $0.974/0.997$, recall 0.828, and precision 0.951. Its FPR mean is sensitive to few negative windows, so we also use a paired moving-block bootstrap on the smallest non-overlapping scale. With block lengths four and eight, the router-minus-fusion FPR differences have 95% intervals $[-0.417, -0.027]$ and $[-0.733, -0.022]$; precision differences are positive. AUROC and AUPRC differences favor the router but include zero because many resamples tie. Gas Sensor Array Drift, screened under a fixed early-to-late batch protocol, produces no accuracy-stable hidden failures and is retained as a negative control (Vergara, 2012). Because candidate streams were screened for endpoint prevalence, this is not a fully prospective data-set trial, although no monitor score or alarm was inspected before freezing the router.

Does generic drift detection solve the task? Table 9 evaluates C2ST-HGB and linear MMD on the same accuracy-stable windows and five-point endpoint across Electricity, Airlines, and Bank Marketing. C2ST-HGB and MMD detect persistent temporal covariate change, but rank harmful versus benign shifted windows substantially worse than the mechanism-aware scores. Their source-null alarms also saturate: C2ST flags every benign window and MMD flags 98.0%. This is not a failure to detect shift; it is evidence that shift magnitude alone does not identify conformal harm.

Small target-label budget. We repeat uniform label acquisition 30 times independently in every Electricity window, using nested budgets of 16, 32, 64, and 128 labels. This produces 87,840 probe evaluations over the same 366 base windows. On the hidden five-point endpoint, the model-assisted 128-label probe obtains group-macro AUROC/AUPRC 0.995/0.956. Its positive-degradation screening rule has mean recall 0.767 and FPR 0.017; the empirical repeat intervals are 0.616–0.936 and 0.007–0.028, respectively. The stricter rule that certifies $G > 0.05$ has recall only 0.272 but FPR 0.0005. At 64 labels, screening recall falls to 0.426 while FPR remains 0.008. Thus labels provide a strong ranking and low-FPR screening fallback, but 128 queries are still insufficient for high-power five-point certification. Appendix C.1 reports the full budget curve and two BRFSS transfer audits.

We next reproduce the TableShift NHANES Childhood Lead task (Gardner et al., 2023). The data contain 27,499 children from ten 1999–2018 NHANES cycles. Following the official split, family poverty-income ratio (PIR) above 1.3 is source and PIR at most 1.3 is target; the PIR-derived domain indicator is excluded from predictors. The label is blood lead at least 3.5 μg/dL. Three models and three seeds generate 270 target batches of 256. Source coverage is 0.901–0.905 across models, versus

Table 9: Strong generic shift baselines on three natural hidden-failure streams. Ranking metrics are dataset-macro mean ± standard deviation over three predictor seeds; alarms use source-only thresholds.

Score	AUROC	AUPRC	Recall	FPR	Precision
C2ST-HGB	0.582 ± 0.018	0.343 ± 0.030	1.000	1.000	0.242
Linear MMD	0.746 ± 0.029	0.529 ± 0.002	1.000	0.980	0.244
Generic drift	0.790 ± 0.024	0.566 ± 0.004	1.000	1.000	0.242
Dual-path fusion	0.965 ± 0.018	0.825 ± 0.025	**0.864**	0.203	0.503
Source router	$\mathbf{0.965 \pm 0.025}$	$\mathbf{0.894 \pm 0.035}$	0.840	**0.146**	**0.640**

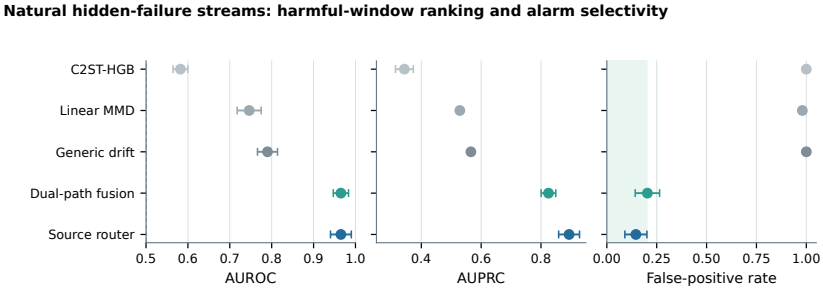


Figure 5: Natural streams: generic tests detect benign change, whereas conformal-risk channels and source-only routing better isolate harmful windows.

0.829–0.840 on target; 221 batches have $G \geq 0.05$. Accuracy also drops by about 5.4 points, so this is a natural accuracy-visible failure stream rather than evidence for the hidden endpoint.

The reversals are informative. Dir-SimCal is strongest on Electricity's accuracy-stable failures, whereas Output-UPE is strongest on both accuracy-visible NHANES failures and accuracy-stable Airlines failures. Bank Marketing shows that source simulations can select between these channels before target labels are observed. Dual-path slightly improves NHANES ranking and halves its FPR, but pays a recall cost; it does not improve hidden-failure ranking on Electricity or Airlines. These results support complementary deployment modes, not universal dominance by one score.

A source-dictionary sensitivity audit is reported in Appendix C. Across 11 variants covering 60–400 simulations, four leave-one-corruption-family-out dictionaries, a restricted severity range, and benign-gap cutoffs from 0.03 to 0.10, hidden-failure recall remains 0.833 and FPR remains 0.046–0.051.

We further stress-test natural shift with three Folktables ACS tasks (Ding et al., 2021). In the geographic audit, California is the source and ten other states are untouched targets; the original grid contains 6,480 evaluations over three models, three seeds, three α levels, and eight target batches. In the temporal audit, California 2014 is the source and the 2015–2018 surveys are target years, yielding another 2,592 evaluations. We add a classifier two-sample test and two scale-free variants—cosine boundary–drift alignment and SCBS divided by total drift. At $\alpha = 0.10$, the geographic audit contains 30 five-point failures and the temporal audit contains 73, but none satisfy the one-point accuracy constraint. Geographic SCBS raises alarms on 99.4% of accuracy-stable windows. Direction-aware ranking improves geographic macro AUROC from 0.521 to 0.571 and annual macro AUROC from 0.564 to 0.570, but Dir-SimCal raises no ACS alarms and recalls none of these accuracy-visible failures. A prespecified pilot that conditions the benign envelope on total drift also produces zero alarms and is not promoted to a full result. These audits supply strong benign-drift controls and expose a transfer limitation of the source intervention dictionary. They do not add evidence for the hidden-failure endpoint.

Table 10: TableShift/NHANES natural domain audit at $\alpha = 0.10$ for $G \geq 0.05$ (270 windows; 221 failures).

Method	AUROC	AUPRC	Recall	FPR	Precision
Dir-SimCal	0.548	0.849	0.014	**0.000**	1.000
Output-UPE	0.910	0.979	**0.724**	0.082	0.976
Dual-path	**0.918**	**0.980**	0.606	0.041	**0.985**

B.8 Computational cost

Across 27 dataset–model cells, offline source profiling takes 5.23 seconds on average and 0.12 seconds at the median; the skew comes from the largest dataset/model combinations. Target drift profiling takes 13.42 ms per batch on average (53.5 μs per sample), while the final risk aggregation costs 1.55 μs. SCBS therefore moves almost all work offline, and its incremental online cost after drift profiling is negligible.

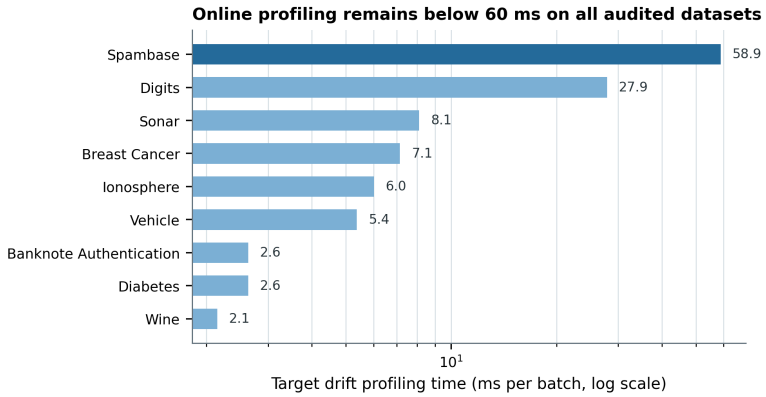


Figure 6: Runtime of unlabeled target drift profiling; SCBS aggregation after this step averages 1.55 μs.

B.9 What the current evidence does not establish

The external, geographic, temporal, Airlines, Bank Marketing, and NHANES studies increase task and shift diversity, but they do not create an arbitrary-shift coverage guarantee. Natural hidden failures occur in three streams. Electricity and Airlines favor different label-free signals, while the frozen Bank Marketing audit confirms that source simulations can route between them. Across the three streams, the router has data-set-macro AUROC/AUPRC $0.965 \pm 0.025/0.894 \pm 0.035$, recall 0.840, FPR 0.146, and precision 0.640; fixed fusion has $0.965 \pm 0.018/0.825 \pm 0.025$, recall 0.864, FPR 0.203, and precision 0.503. Thus routing improves selectivity rather than uniformly dominating fusion. NHANES adds an accuracy-visible regime. Uniform target-label probing addresses otherwise unidentified conditional change only by changing the information setting. The external hidden endpoint also has few valid datasets at several α/batch combinations, and three natural tasks remain too few for a universal transfer claim. We do not use correlated window counts as a substitute for independent task diversity.

C Additional Experimental Details

Source-only temperature selection. For each dataset/model/seed cell, an auxiliary source corruption bank selects the temperature multiplier without target outcomes. This avoids target-label tuning but is more variable than fixing a global multiplier; we therefore retain multiplier 1 for the formal main comparison and report the source-selected variant as a sensitivity analysis.

Table 11: Dir-SimCal source-dictionary sensitivity on hidden Electricity failures.

Variant family	Settings	Recall	FPR	Precision	Min. benign
Simulation-bank size	60	0.833	0.051	0.559	54
	100	0.833	0.051	0.559	88
	200	0.833	0.048	0.567	179
	400	0.833	0.048	0.567	358
Leave-one-family-out	four exclusions	0.833	0.048–0.051	0.559–0.578	129
Severity dictionary	≤ 1.0 only	0.833	0.051	0.559	140
Benign-gap cutoff	0.03–0.10	0.833	0.046–0.048	0.567–0.585	173

Table 12: Model-assisted label-probe sensitivity on hidden Electricity failures at $\alpha = 0.10$ (30 sampling repetitions).

Labels	AUROC	AUPRC	Screen rec.	Screen FPR	Cert. rec.	Cert. FPR
16	0.932	0.704	0.168	0.009	0.073	0.0021
32	0.967	0.798	0.265	0.008	0.083	0.0013
64	0.986	0.895	0.426	0.008	0.152	0.0012
128	0.995	0.956	0.767	0.017	0.272	0.0005

Failure prevalence. Of 57,600 total scenarios, 25,183 have an accuracy drop at most 0.01. Within this subset, 2,504 (9.95%) have a coverage gap of at least 0.03, 981 (3.90%) at least 0.05, and 185 (0.73%) at least 0.10. Dataset-macro AUPRC is interpreted relative to these low and dataset-varying base rates.

Source-simulation dictionary sensitivity. We rerun the Electricity $\alpha = 0.10$ audit with a shared bank of 400 source pseudo-deployments and vary one calibration choice at a time. Metrics are macro-averaged over nine model–window-size cells on the hidden five-point endpoint (246 windows; 17 failures). Table 11 shows that recall is unchanged across all 11 variants and FPR remains between 0.046 and 0.051. Increasing the bank beyond 200 provides no measurable benefit; removing any one corruption family also preserves the operating point. This supports stability within the tested dictionary, but does not imply that these source interventions span every natural domain shift.

C.1 Finite-label probe sensitivity

The Electricity probe uses the same chronological splits, three models, and window sizes as the label-free natural audit. Each of 30 repetitions draws an independent uniform permutation within every target window; budgets are nested prefixes, so comparisons across budgets are paired. Metrics are first computed within each model–window cell, macro-averaged within a repetition, and then summarized over repetitions. Table 12 separates screening for any positive degradation from one-sided 95% certification that the gap exceeds five points.

The regression adjustment is helpful but modest relative to uniform labeling itself. At 128 labels, the naive estimator obtains AUROC/AUPRC 0.993/0.938 and screening recall/FPR 0.701/0.015; model assistance changes these to 0.995/0.956 and 0.767/0.017. We therefore present the adjustment as a variance-reduction device rather than the source of identifiability.

We also apply one nested probe draw per seed/window to two TableShift-style BRFSS 2021 domain shifts. With 32 labels, model-assisted AUROC/AUPRC and screening recall/FPR are 0.778/0.520 and 0.465/0.066 for diabetes/race, versus 0.723/0.607 and 0.200/0.079 for blood-pressure/BMI. At 128 labels the recall rises above 0.93 on both tasks, but FPR also exceeds 0.41. Neither BRFSS task contains a five-point failure satisfying the one-point accuracy constraint, so they test accuracy-visible transfer rather than supply a second hidden failure stream.

C.2 Source-simulation channel routing and frozen confirmation

The two natural streams expose complementary failure mechanisms: the directional score is stronger on Electricity, whereas Output-UPE is much stronger on Airlines. We therefore study a source-only

router that selects one channel within each model–α–window-size cell. Each channel is divided by its 95th percentile over benign source pseudo-deployments. When at least five harmful source pseudo-deployments are available, the router selects the channel with the larger source-simulation harmful-versus-benign AUROC; it otherwise falls back to Output-UPE. Neither scoring nor routing uses outcomes from natural target windows.

Across three seeds, the router selects the directional channel in 7, 6, and 5 of nine Electricity cells, but selects Output-UPE in 9, 8, and 9 of nine Airlines cells. These observations fixed the rule. On score-blind Bank Marketing, it then selects the directional channel for all three logistic-regression scales and Output-UPE for all six tree-model scales in every seed. Table 13 macro-averages all three datasets within each seed and then reports the mean and standard deviation across seeds. The router has the best AUPRC and improves the fusion operating point's false-positive rate and precision, with a small recall cost and essentially equal AUROC. Thus it improves selectivity rather than uniformly dominating fusion.

Table 13: Three-stream channel routing on accuracy-stable natural windows with five-point coverage failures.

Score	AUROC	AUPRC	Recall	FPR	Precision
Dir-SimCal	0.883 ± 0.026	0.811 ± 0.009	0.652	0.120	0.535
Output-UPE	0.808 ± 0.028	0.688 ± 0.031	0.770	0.173	0.481
Dual-path fusion	0.965 ± 0.018	0.825 ± 0.025	**0.864**	0.203	0.503
Source router	$\mathbf{0.965 \pm 0.025}$	$\mathbf{0.894 \pm 0.035}$	0.840	**0.146**	**0.640**

The Electricity/Airlines comparison remains developmental. Bank Marketing was evaluated only after the routing rule was frozen, and its candidate screen used endpoint prevalence but not monitor scores. We therefore treat it as score-blind frozen evidence rather than a fully prospective data-set trial, and the three-stream aggregate as a small-task result rather than a universal guarantee. The paired moving-block analysis resamples the full 256-record timeline before applying the accuracy-stable filter and uses the same temporal indices for all models and seeds, avoiding pseudo-replication from overlapping windows.