REFINE, DON'T MERGE: CERTIFIED REPAIR OF BLIND CONFORMAL DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential conformal detectors can control false alarms yet miss changes that preserve their score distribution. This matters for score-based monitoring, including prospective monitoring of language-model agents. We formalize these *rank-fiber collisions* and introduce CoFi-E, which refines inherited score groups and independently certifies each retained split's information gain. We prove information retention and a geometric rate for population greedy refinement under a weak-witness condition. Freezing the repair preserves runtime false-alarm control under conditional exchangeability. In a controlled joint collision, CoFi-E detects 84.1% of streams versus 1.4% for marginal merging. Experiments span image, time-series, and intrusion data. Across 290 matched tasks, selected partitions reduce restricted delay by 34.8% against the inherited detector and 18.8% against a fixed grid; audit-gated prefixes yield 26.7% and 24.6%, respectively. These results support certified repair of specified detector blind spots.

1 INTRODUCTION

A sequential detector can satisfy its false-alarm guarantee and still miss every change that its score erases. Conformal detectors compare incoming scores with normal reference scores and accumulate evidence from their *ranks*, or relative positions (Volkhonskiy et al., 2017; Vovk et al., 2021). Under conditional exchangeability, this evidence bounds the probability of ever raising a false alarm (Vovk et al., 2003; Ramdas et al., 2023; Shin et al., 2024). But if a change preserves the score distribution, the rank stream can have exactly its normal law. No evidence rule using only that interface can expose the change (Figure 1a–b).

This limitation is relevant to monitoring language-model agents. AI-control protocols use model-based monitors to assess potentially harmful behavior (Greenblatt et al., 2024), while conformal methods provide statistical guarantees for language-model outputs (Quach et al., 2024). Consider a monitor that groups tool-use records by a risk score. If a change shifts probability toward harmful actions within each group while preserving group frequencies, a conformal detector receiving those groups remains blind. This is a possible failure mode of the chosen representation, even when calibration is correct. It motivates the repair problem studied here; our experiments test its mechanism on image, time-series, and intrusion data, and do not evaluate LLM agents or certify their safety.

The repair question is: how can a detector expose a newly identified blind spot while retaining what it already distinguishes? Replacing the score can discard earlier distinctions (Cherubin et al., 2018; Vovk et al., 2021). Combining old and new detectors preserves their separate views (Eliades & Papadopoulos, 2023), but a fixed weighted average of their evidence processes can miss a change that preserves both marginal score laws and alters only their joint law (Proposition 1). Refinement addresses this obstruction by dividing existing score groups, retaining the inherited distinctions and exposing variation within them.

We introduce **CoFi-E** (Collision-guided Fiber-refining Conformal E-detection) to make this repair statistically auditable. Examples of target changes guide candidate divisions inside the current groups; separate data select a complete refinement path. Independent audit counts then certify a positive information gain for each retained division. The audit controls erroneous gain claims, while a separate runtime guarantee controls false alarms. Nestedness preserves population score information for every alternative; a positive certificate exposes additional information for the targeted alternative. The accepted representation is frozen before reference data are observed (Figure 1d–f).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

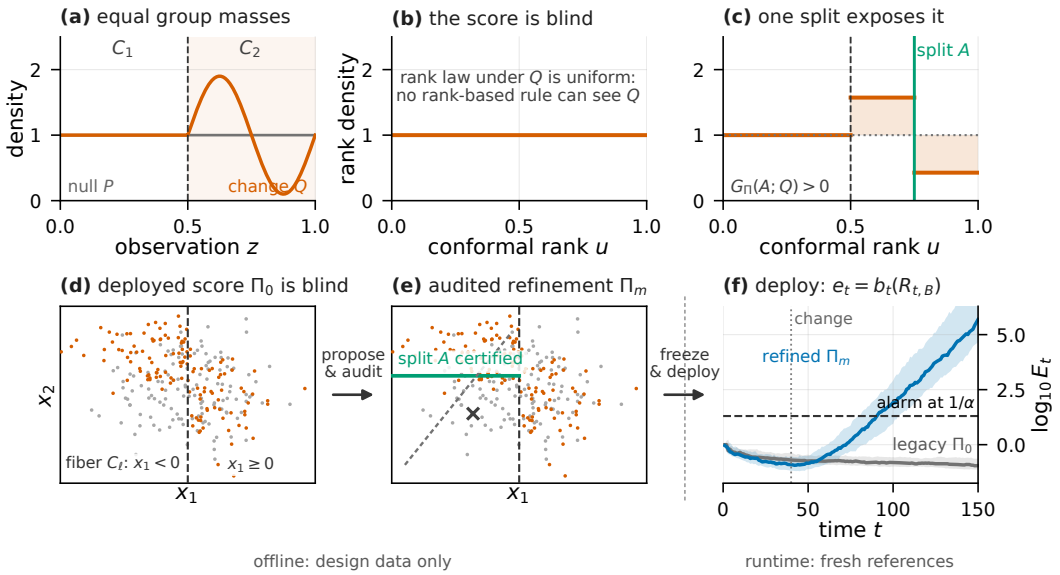


Figure 1: (a–c) A rank-fiber collision. (a) The alternative changes observations inside score group C_2 while preserving every group mass. (b) Its rank law is uniform, so this interface is blind. (c) Splitting C_2 exposes the hidden change as a nonuniform population rank law. (d–f) The three stages of CoFi-E on the running example of Section 2. The crossed proposal fails its audit, so the certified prefix Π_m retains only the division that was certified. Panels (d–e) are schematic; (f) simulates 120 streams on a fixed refined map with $B = 63$ (median and interquartile range).

Conditional partition refinement already appears in multicalibration and importance weighting (Hebert-Johnson et al., 2018; Gopalan et al., 2022). CoFi-E studies the additional requirements imposed by repairing an inherited conformal interface: permanent nesting, independent gain certificates, and runtime validity after repair. This formulation connects a local discrepancy to both the information recovered and the distinctions retained (Section 7). We make three contributions:

1. We formalize *rank-fiber collisions* and characterize a limitation of fixed-weight marginal process mixtures that a joint refinement can overcome (Section 3).
2. We give a count-based repair algorithm and prove information retention, simultaneous gain certification, and a population refinement rate under a weak-witness condition (Sections 4–5).
3. We test information recovery, sequential detection, and search choices separately. Controlled collision detection rises from 1.4% for merging to 84.1% for CoFi-E; 290 matched tasks favor targeted splitting over fixed and random partitions. Image, time-series, and intrusion studies evaluate targeted repair and retention (Section 6).

2 What the Detector Sees at Runtime

We specify the inherited score interface and the sampling conditions for runtime validity.

Normal and target data. Let P denote normal behavior on an observation space $\mathcal{Z}$, and let $Q_1, \ldots, Q_J$ denote the J target changes that the detector is designed to find. We assume $Q_j \ll P$, so a target change does not place probability on an event that is impossible under normal behavior. All representation choices use design data only, before runtime reference data are observed.

Score groups. The deployed score puts observations into L groups, $\Pi = \{C_1, \ldots, C_L\}$. We call C_ℓ a *score group*; it is also a fiber of the score because the detector does not distinguish observations inside it. The categorical score is $s_\Pi(z) = \ell$ when $z \in C_\ell$. A refinement $\Pi' \succeq \Pi$ divides existing score groups but never combines them. For a target change Q, write

$$p_\ell = P(C_\ell), \qquad q_\ell = Q(C_\ell).$$

These are the normal and target probabilities of score group ℓ. We assume $p_\ell > 0$ and fix the group ordering before calibration.

Running example. Suppose a legacy score groups observations only by whether a first feature is low or high. A target change can alter a second feature within the low group while leaving the probability of the low and high groups unchanged. The legacy score is then blind to the change; CoFi-E can divide the low group using the second feature without removing the original low–high distinction.

Population rank. For a new observation Z, compare its score with an independent normal observation $Z' \sim P$. The following randomized rank handles ties fairly:

$$R_\Pi = P\{s_\Pi(Z') < s_\Pi(Z)\} + U_{\text{tie}}\, P\{s_\Pi(Z') = s_\Pi(Z)\}.$$

Here U_{tie} is an independent uniform random number. Equivalently, score group ℓ maps uniformly to a disjoint interval $\mathcal{I}_\ell \subset [0,1]$ of length p_ℓ. Under normal behavior, R_Π is uniform on $[0,1]$.

Finite-reference rank and false-alarm control. At time t, Z_t is the monitored observation and $X_{t1:B}$ are B fresh normal reference observations. Let $S_t = s_\Pi(Z_t)$ and $S_{ti} = s_\Pi(X_{ti})$ for $i = 1, \dots, B$. With an independent $U_t \sim \text{Unif}(0,1)$, the finite-reference version of the rank is

$$R_{t,B} = \frac{\sum_{i=1}^{B} \mathbf{1}\{S_{ti} < S_t\} + U_t\big(1 + \sum_{i=1}^{B} \mathbf{1}\{S_{ti} = S_t\}\big)}{B+1}. \tag{1}$$

The first sum counts lower reference scores; the second randomizes ties. Under normal behavior, assume the score bundle is exchangeable conditional on the past and frozen design data. Then $R_{t,B}$ is conditionally uniform (Vovk et al., 2003).

An evidence function $b_t : [0,1] \to [0,\infty)$ is chosen using only information available before time t and satisfies $\int_0^1 b_t(u)\, du \le 1$. Its output $e_t = b_t(R_{t,B})$ is one-step evidence, and $E_t = \prod_{k=1}^{t} e_k$ accumulates evidence over time. Under normal behavior, E_t is a nonnegative supermartingale, so by Ville's inequality (Ville, 1939) the first alarm $\tau_\alpha = \inf\{t : E_t \ge 1/\alpha\}$ obeys

$$\mathbb{P}_P(\tau_\alpha < \infty) \le \alpha. \tag{2}$$

This is the false-alarm guarantee. CoFi-E changes the score representation before monitoring begins; it does not change this runtime argument. Population rank information and finite-reference evidence are different quantities. Proposition 3 gives the finite-reference transfer result.

3 When a Score Becomes Blind

A score can hide a real change when the change occurs inside its score groups. For a score s, the KL chain rule separates the information visible in the score from the information hidden within groups:

$$D_{\text{KL}}(Q\|P) = D_{\text{KL}}(Q^s\|P^s) + \mathbb{E}_Q[D_{\text{KL}}\{Q(\cdot \mid s)\|P(\cdot \mid s)\}]\,. \tag{3}$$

The first term is information the score retains. The second term is information it erases. Here Q^s and P^s are the target and normal score distributions; $Q(\cdot \mid s)$ and $P(\cdot \mid s)$ describe observations that receive the same score (Cover & Thomas, 2006).

Definition 1 (Rank-fiber collision attack)**.** For a score representation Π, a *rank-fiber collision* is a target distribution $Q \neq P$ with $Q^\Pi = P^\Pi$. In words, it changes the observations but leaves the distribution of score groups unchanged.

Every rank-fiber collision has the same population rank distribution as normal behavior. With independent monitored observations and fresh normal reference bundles, the entire rank stream has the same law under P and Q. Thus every detector that uses only this score is blind, not merely one chosen threshold or evidence function (Figure 1a–b). Appendix D discusses extensions to context-aware and multiple-score interfaces.

Algorithm 1: Certified CoFi-E

Input: Frozen score groups Π_0; disjoint proposal, selection, and audit data; proposal budget N_{prop}; audit level δ and minimum normal support.

Output: Certified partition Π_m and frozen runtime evidence function b.

1 Build candidate divisions from the proposal data, restricting each division to one current score group.
2 Use the selection data to construct the complete greedy path $\Pi_0, \ldots, \Pi_{N_{\mathrm{prop}}}$.
3 Audit every proposed division: compute $\underline{G}_r$ and check its new score groups for declared normal support.
4 Keep the longest prefix certified by positive lower bounds and normal support.
5 Freeze Π_m and fit b on selection data in its score groups.

3.1 The limit of marginal process merging

A legacy score s_0 and a new score s_1 can each miss a change in how the two scores occur together. The following result concerns a fixed weighted average of their evidence processes.

Proposition 1 (Merge blindness)**.** *Suppose queries are iid from Q, and each fresh iid normal reference bundle is independent of the queries and past bundles. Assume Q preserves the normal score distribution for each frozen s_j, $j \in \{0, 1\}$. Each normalized predictable evidence process $M_j(t)$ then remains a nonnegative supermartingale under Q. For fixed weights $w_j \geq 0$ with $\sum_j w_j = 1$, their average $M(t) = \sum_j w_j M_j(t)$ satisfies $\mathbb{P}_Q\{\sup_t M(t) \geq 1/\alpha\} \leq \alpha$.*

The displayed bound says that averaging cannot reliably expose a change that each separate score misses. The proof uses linearity and Ville's inequality (Appendix F). A joint refinement can retain the information that the marginal average does not, as shown next.

Lemma 1 (Refinement separates joint changes)**.** *If score groups Π refine both score representations Π_0 and Π_1, then*

$$D_{\mathrm{KL}}(Q^{R_\Pi} \| U) \geq D_{\mathrm{KL}}\left(Q^{(\Pi_0,\Pi_1)} \middle\| P^{(\Pi_0,\Pi_1)}\right).$$

The right side measures the change visible only when the two scores are read together. Equality holds for their product representation, so a joint refinement retains this information while a marginal average does not.

The product score is a coarser view of Π, so data processing and Lemma 2 give the inequality. A fully supported $L_0 \times L_1$ score table has $(L_0 - 1)(L_1 - 1)$ ways to change its joint pattern while keeping both separate score distributions fixed. Joint score groups observe those changes; separate marginal detectors do not.

4 How CoFi-E Builds a Repair

4.1 Propose, select, audit, and freeze

CoFi-E starts from the deployed score groups Π_0, with the one-group representation as a special case. Three disjoint design datasets propose divisions, select a complete path, and audit that fixed path. Neither audit nor runtime observations enter proposal or selection.

Proposal data build a logistic-halfspace dictionary with axis-aligned fallbacks (Lopez-Paz & Oquab, 2017; Menon & Ong, 2016), restricting each candidate division to one current score group. Selection ranks candidates by estimated gain, and the independent audit retains a certified prefix (Appendix B.6).

For a declared target mixture, the fitted evidence function may use smoothed group probabilities $\hat{p}, \hat{q}_{1:J}$ and a likelihood-ratio component mixed with the constant-one function.

The population rate in Theorem 2 requires additional selection and audit-power assumptions for Algorithm 1. CoFi-E searches supplied target changes; it does not generate new attacks during refinement.

4.2 MEASURING A PROPOSED DIVISION

To target a blind spot, the split criterion must measure information still hidden inside a group. Subtracting two learned KL estimates can be unreliable (Song & Ermon, 2020). We use a Pearson decomposition whose local gain can be computed from counts. The information already represented is

$$\mathrm{Vis}_\Pi(Q) = \sum_{\ell=1}^{L} \frac{q_\ell^2}{p_\ell} - 1. \tag{4}$$

The residual $\mathrm{Hid}_\Pi(Q)$ measures information hidden within groups. Theorem 1 shows that each split gain equals the reduction in this residual, giving the audit a direct measure of repair.

Theorem 1 (Fiber projection and characteristic completeness)**.** *Let $r = dQ/dP$ be the target-to-normal likelihood ratio and assume $r \in L^2(P)$. Then*

$$\mathrm{Hid}_\Pi(Q) = \chi^2(Q\|P) - \mathrm{Vis}_\Pi(Q), \tag{5}$$

$$\mathrm{Hid}_\Pi(Q) - \mathrm{Hid}_{\Pi'}(Q) = \mathrm{Vis}_{\Pi'}(Q) - \mathrm{Vis}_\Pi(Q) \geq 0 \quad (\Pi' \succeq \Pi). \tag{6}$$

The first line defines hidden information as full information minus visible information. The second says that a division moves information from hidden to visible. If A is a proposed child group inside parent group C, with $p_C = P(C)$, $q_C = Q(C)$, $p_A = P(A)$, $q_A = Q(A)$ and $0 < p_A < p_C$, its gain is

$$G_\Pi(A; Q) = \frac{(p_C q_A - q_C p_A)^2}{p_C p_A (p_C - p_A)}. \tag{7}$$

The gain is positive exactly when normal and target data place different conditional mass in the proposed child group. Moreover, $\mathrm{Hid}_\Pi(Q) = 0$ iff the two conditional distributions agree inside every Q-positive score group. It is enough to test a measure-determining class of candidate divisions; affine halfspaces are one such class on $\mathbb{R}^d$.

A measure-determining class identifies a distribution from its group probabilities; halfspaces suffice in d dimensions (Cramér & Wold, 1936). Equation (7) uses only four probabilities, so independent counts can certify it exactly. Proofs are in Appendix F.

4.3 EXACT AUDIT CERTIFICATE

For a proposed child group A inside parent group C, the audit asks whether its conditional probability differs between normal and target data. Write $u = P(A \mid C)$ and $v = Q(A \mid C)$. Equation (7) becomes $G_\Pi(A; Q) = (q_C^2/p_C)(v-u)^2/\{u(1-u)\}$. Let I_p, I_q, I_u, I_v be simultaneous Clopper–Pearson intervals for p_C, q_C, u, v, respectively (Clopper & Pearson, 1934), and let $d(I_u, I_v) = \inf_{x \in I_u, y \in I_v} |x - y|$. Then

$$\underline{G} = 4\frac{\underline{q}_C^2}{\overline{p}_C} d(I_u, I_v)^2 \leq G_\Pi(A; Q). \tag{8}$$

The lower bound uses the most conservative values allowed by the four confidence intervals. It is positive only when the data certify a difference inside the old score group.

For N_{prop} proposals and total audit error level δ, give each of the four intervals error $\delta/(4N_{\mathrm{prop}})$. A union bound then gives simultaneous coverage at least $1 - \delta$, including the fixed selection path. An absent conditional count yields the conservative interval $[0, 1]$ and therefore a zero lower bound.

Remark 1 (What the certificate means)**.** $\underline{G} > 0$ certifies that one particular alternative differs from the null inside one current score group and that the division exposes positive Pearson information. It does not certify useful ordering for an undeclared alternative. Appendix E states the additional coverage condition.

4.4 ADAPTIVE EVIDENCE ACCUMULATION

After the representation is frozen, CoFi-E can use a Krichevsky–Trofimov (KT) evidence function (Krichevsky & Trofimov, 1981; Orabona & Pál, 2016). Let $n_{t-1,\ell}$ count earlier arrivals in score group

ℓ, and let $K_{B,\hat{p},\ell}$ be the finite-reference rank density for that group under the frozen normal-group estimate $\hat{p}$. The update is

$$\widehat{q}_{t,\ell} = \frac{n_{t-1,\ell} + 1/2}{t - 1 + L/2}, \qquad b_t(u) = (1-\eta) + \eta \sum_{\ell=1}^{L} \widehat{q}_{t,\ell} K_{B,\hat{p},\ell}(u).$$

The current count is added after its rank is scored. Thus b_t is fixed before each rank and integrates to one for every history, even when $\hat{p}$ is imperfect. Conditional rank uniformity supplies false-alarm control; estimation affects power. Appendix B.5 gives details and Appendix F proves validity.

4.5 Replace, merge, or refine

Replacement learns a new score and can discard distinctions made by the old one. Merging keeps separate rank processes and inherits Proposition 1. Refinement keeps all old score groups, certifies divisions inside them, and can add the adaptive component above. CoFi-E+ is a fixed mixture of legacy, standalone, declared-refined, and adaptive-refined evidence processes. If component i has weight $w_i > 0$, the mixture retains at least fraction w_i of its evidence at every time. It therefore loses at most $\log(1/w_i)$ in log evidence while retaining access to joint changes (Vovk & Wang, 2021).

5 Why Refinement Works

5.1 Dividing score groups cannot erase information

The audit certifies a gain for a specified target. Nestedness supplies a complementary guarantee: information retained by the inherited partition cannot decrease for any alternative. The results below establish this retention property and the conditions for progressively reducing hidden information.

Lemma 2 (Partition-rank identity)**.** *For every $Q \ll P$, the density of the population rank under Q is $w_{Q,\Pi}(u) = q_\ell / p_\ell$ on score-group interval $\mathcal{I}_\ell$. Therefore*

$$D_{\mathrm{KL}}(Q^{R_\Pi} \| U) = D_{\mathrm{KL}}(Q^{\Pi} \| P^{\Pi}). \tag{9}$$

The left side is information available from the rank; the right side is information available from the score groups. If Π' divides score groups of Π, then

$$D_{\mathrm{KL}}(Q^{\Pi'} \| P^{\Pi'}) = D_{\mathrm{KL}}(Q^{\Pi} \| P^{\Pi}) + \sum_{C \in \Pi} Q(C) d_C, \qquad d_C = D_{\mathrm{KL}}(Q(\Pi' \mid C) \| P(\Pi' \mid C)). \tag{10}$$

Here d_C is the difference between normal and target child-group probabilities within parent group C. Every term added by a division is nonnegative, so no nested refinement reduces rank information. Once the refined score is frozen, the runtime protocol in Section 2 still satisfies (2)*.*

Corollary 1 (Certified inherited-interface retention)**.** *If Algorithm 1 starts from a frozen finite Π_0 and returns $\Pi_m \succeq \Pi_0$, then for every $Q \ll P$,*

$$D_{\mathrm{KL}}(Q^{R_{\Pi_m}} \| U) \geq D_{\mathrm{KL}}(Q^{R_{\Pi_0}} \| U)$$

deterministically. On the simultaneous audit event, every declared target that appears in at least one retained split has a strict positive increase in population rank KL relative to Π_0.

A positive certificate makes one conditional-KL increment strictly positive; nestedness makes all others nonnegative. This is population-information retention, not guaranteed monotonic finite-reference evidence. The proof is in Appendix B.

5.2 Greedy refinement and its rate

At round k, CoFi-E chooses the target change and candidate division with the largest gain. For a finite candidate dictionary, this is

$$(j_k, A_k) \in \operatorname*{arg\,max}_{j \leq J,\ A \in \mathcal{A}_{\Pi_k}} G_{\Pi_k}(A; Q_j), \qquad \Pi_{k+1} = \Pi_k \vee A_k.$$

Here $\mathcal{A}_{\Pi_k}$ is the available set of divisions, j_k identifies the selected target change, and $\Pi_k \vee A_k$ divides the parent group into A_k and its complement.

Population refinement rate. Suppose the candidate dictionary always contains a division that exposes at least a fixed fraction γ of the hidden information for each remaining target change. Then the exact population greedy rule reduces total hidden information geometrically, by a factor no larger than $1 - \gamma/J$ at each division. This is a population result: Algorithm 1 uses estimated gains and an independent audit, so it needs additional selection and audit-power assumptions to inherit this rate. The complete theorem, its additive-error version, and finite-alphabet consequences appear in Appendix B.1.

6 EXPERIMENTS

We test information recovery, sequential detection, and the split criterion. Controlled collisions test the mechanism; image, time-series, and intrusion data test repair and transfer; matched ablations fix the runtime rule and cell budget. Protocols and correction records are in Appendix G. Detection is the fraction of all streams whose first alarm occurs at or after onset and within the horizon. Early alarms remain in the denominator and are reported separately or scored as failures.

6.1 RECOVERING HIDDEN INFORMATION WHILE RETAINING EARLIER DISTINCTIONS

On four equally likely points, an inherited score records the horizontal coordinate. A vertical mass shift is hidden; a horizontal shift is visible. With identical samples, audit rules, and a four-group cap, CoFi-E certifies both required splits in each of five seeds. Repair raises the hidden family's population rank KL from zero to 0.1927 nats while retaining 0.1927 nats for the earlier family. Unrestricted refitting recovers the hidden family but loses the earlier distinction. Exact $B = 63$ evaluation of a mixture of legacy and newly fitted evidence gives positive log growth for both families (0.0972, 0.0600) after repair (Table 3, Appendix G.1).

A second alternative preserves both marginal score laws but changes their joint table. Its population joint KL is 0.08228 nats and finite-bundle rank KL is 0.06200. At onset 25, plain CoFi-E detects 0.8414 of streams, versus 0.0138 for process merging; null crossing rates are 0.0310 and 0.0378 (Figure 2a). Plain CoFi-E equally mixes declared-refined and adaptive-refined processes. CoFi-E+ obtains 0.8264 detection and 0.0344 null crossing with weights $(0.10, 0.10, 0.40, 0.40)$ on legacy, standalone, declared-refined, and adaptive-refined components. The earlier equal-weight CoFi-E+ run and subsequent weight change are disclosed in Appendix G.1. This construction isolates joint information and adaptive betting. The learned partition equals the fixed product partition, so a product comparator with the same evidence function is identical to plain CoFi-E. The matched study below evaluates partition search separately.

6.2 REPAIR ACROSS IMAGE, TIME-SERIES, AND INTRUSION DATA

Table 1 evaluates seven specified blind spots on four image and time-series datasets (Bergmann et al., 2019; Dau et al., 2019; Bagnall et al., 2018), using reserved observations. At onset 25, CoFi-E+ combines 97.9% target detection with 97.4% earlier-family retention, versus $97.8\%/25.3\%$ for logistic replacement and $97.7\%/95.7\%$ for merging. CoFi-E+ includes the merged components. At onset 100, CoFi-E and CoFi-E+ detect $14.1\%/13.3\%$, versus 0.4% for merging. Appendix G.2 gives the comparison's interpretation and the separate thirteen-artifact collision evaluation (Table 4).

The corrected UNSW-NB15 (Moustafa & Slay, 2015) replay certifies $102/160$ proposed splits. At onset 25, the adaptive refined component detects 0.756 of streams, versus 0.424 for marginal merging. Appendix G.2 reports the full evidence-rule comparison, and Appendix G.3 documents the corrected replay. The matched study below holds the evidence rule fixed to evaluate partition choice.

6.3 ABLATING THE SPLIT CRITERION AT MATCHED BUDGETS

The primary panel comprises UNSW-NB15, ElectricDevices, FordA, and FaceDetection: 290 dataset-family-budget-seed tasks, five seeds, $K \in \{8, 16\}$ cells, $B = 63$, and 1,000 streams per condition. Six rules share the inherited four-cell interface: unchanged legacy, fixed joint grid, local quantiles, random cuts, supervised Gini search, and CoFi-E Pearson search. Gini search splits on label impurity reduction rather than the Pearson gain of Eq. (7) (Breiman et al., 1984). The search rules share the candidate dictionary and allowance. The unchanged legacy control retains four cells. All rules are

Table 1: Targeted repair on image and time-series features (35 task–seed records). Mean percentages; $B = 63$, $\alpha = 0.05$, horizon 150. Retention measures earlier-family detection at onset 25; null counts every alarm. Bold marks detection/retention column maxima, not statistical significance.

	Target detection (↑)			Retention (↑)	Null
Method	Onset 25	Onset 50	Onset 100	Onset 25	Alarm
Isolation Forest (legacy)	33.0	16.4	0.4	93.6	3.6
Logistic replacement	97.8	96.2	0.0	25.3	2.4
HGB replacement	83.8	81.2	0.0	36.0	1.8
Marginal merging	97.7	95.6	0.4	95.7	3.2
Fixed product	84.5	58.8	4.5	70.8	3.5
CoFi-E (ours)	85.5	72.1	**14.1**	97.1	3.5
CoFi-E+ (ours)	**97.9**	**96.4**	13.3	**97.4**	3.5

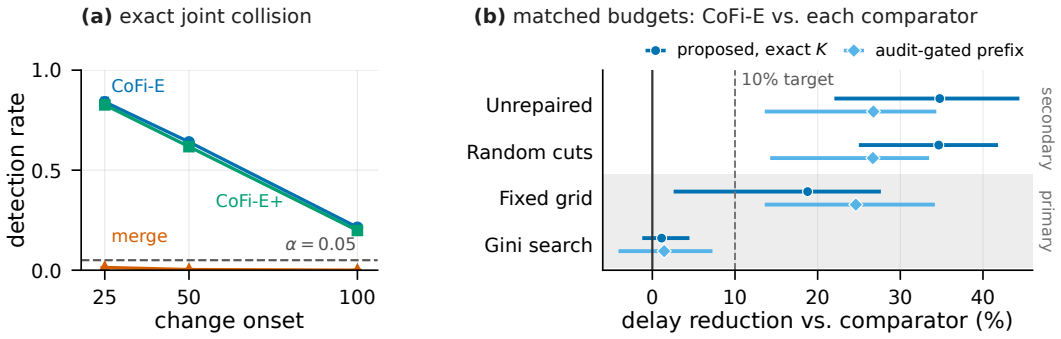


Figure 2: (a) Exact joint collision: detection rate at each change onset (mean and range over five seeds). A product partition with the same evidence rule equals CoFi-E here. (b) Matched cell budgets and adaptive evidence: relative reduction in restricted delay for CoFi-E versus each comparator, with 97.5% paired dataset-cluster bootstrap intervals (values in Table 5). Fixed grid and Gini are the preregistered primary contrasts; the dashed line is the frozen 10% success threshold.

evaluated with declared and adaptive evidence functions; Table 5 holds the adaptive KT evidence function fixed to isolate partition choice.

Restricted delay is alarm time minus onset plus one, with a loss of 151 for a miss or a pre-onset first alarm. Each stream has 150 post-change steps; onsets are 25, 50, and 100. We average over streams, seeds, budgets, and onsets, then families within datasets and datasets equally. The proposed *exact-K* view evaluates all selected splits; the *audit-gated* view evaluates only the certified prefix and can have fewer cells. All 290 tasks are feasible. Intervals use 20,000 paired bootstrap draws that resample whole datasets rather than single tasks (Efron & Tibshirani, 1993), with nominal 97.5% coverage per contrast, giving a Bonferroni pair within each view. With only four dataset clusters, their coverage is approximate.

At exact K, CoFi-E reduces restricted delay by 34.8% against legacy and 18.8% against the fixed grid; audit-gated reductions are 26.7% and 24.6% (Figure 2b). Random cuts improve on legacy by only 0.2% and 0.1%, with intervals spanning zero; local quantiles perform like the fixed grid. These controls support targeting within-group discrepancies over simply adding groups. Figure 2b displays all four contrasts. Appendix G.4.1 reports the Pearson–Gini comparison, the registered success criterion and its outcome, accepted partition sizes, and the scope of sensitivity evaluation.

6.4 Additional Evaluation

Appendix H reports the complete 605-fold unseen-family evaluation on 13 image and time-series datasets (Table 9). Appendix G.4 gives repeated-repair and approximate-collision studies. These evaluations accompany the targeted-repair comparisons above; the family-transfer guarantee and its coverage premise are in Appendix E.

7 Related Work

Sequential conformal detection. Conformal test martingales and e-detectors turn ranks into anytime-valid change detectors (Vovk et al., 2003; Volkhonskiy et al., 2017; Vovk et al., 2021; Shin et al., 2024). Later work reweights, recalibrates, or adapts this evidence under shift (Prinster et al., 2025; Shaer et al., 2026; Gibbs & Candès, 2021). These methods take the score as given; CoFi-E changes the score and leaves the runtime validity argument unchanged.

Learned scores, tests, and partitions. Learned conformity scores improve efficiency or ranking (Cherubin et al., 2018; Stutz et al., 2022; Kiyani et al., 2024). Classifier and kernel two-sample tests learn global discrepancies (Lopez-Paz & Oquab, 2017; Liu et al., 2020; Shekhar & Ramdas, 2024; Podkopaev & Ramdas, 2023), and histogram or tree methods localize them (Boracchi et al., 2018; Zhou et al., 2023). Multicalibrated refinement searches for conditional discrepancies inside current cells (Hebert-Johnson et al., 2018; Gopalan et al., 2022). Building on this principle, CoFi-E imposes permanent retention of an inherited partition, certifies every retained split with independent counts, and connects the result to a valid rank e-process.

Combining detectors and monitoring AI systems. Averaging e-values preserves validity (Vovk & Wang, 2021), and ensembles of conformal martingales use such combinations to track drift (Eliades & Papadopoulos, 2023). Proposition 1 shows that fixed-weight marginal mixtures still inherit a joint blind spot. AI-control protocols rely on model-based monitors (Greenblatt et al., 2024), and conformal methods give guarantees for language-model outputs (Quach et al., 2024). Rank-fiber collisions describe a representation-level failure mode that such score-based monitors can share. Appendix C gives details.

8 Scope

The repair guarantees concern population information in a declared score interface, with audit and runtime validity under their respective sampling assumptions. The refinement rate requires a weak witness, and family transfer requires score-law coverage. Appendix D details deployment requirements, empirical qualifications, and limitations.

9 Conclusion

Rank-fiber collisions separate false-alarm validity from the ability to observe a change. CoFi-E addresses this repair problem by exposing certified within-group differences while retaining the inherited score's population information. Controlled experiments demonstrate recovery without forgetting; matched comparisons support targeted splitting over fixed and random groups. Image, time-series, and intrusion studies evaluate repair of specified changes. Next steps include fixed-calibration runtime rules, conditional guarantees for history-dependent streams such as agent trajectories, step-down audits, and neural or kernel witnesses. These guarantees make the requirements for extending repair to richer interfaces, including language-model agents, explicit.

AI USE STATEMENT

Generative AI assisted coding, editing, and literature organization. Authors verified the code, citations, claims, and final text; experimental values come from the frozen scripts and independent verifiers.

ETHICS STATEMENT

RFCA shows that exact calibration does not imply an information-preserving score, and released transports should be interpreted at their declared interfaces. Experiments use public benchmark datasets only; no human subjects or identifying data are involved.

REPRODUCIBILITY STATEMENT

Code, frozen protocols, result JSONs, figures, and independent verifiers are released at `https://anonymous.4open.science/r/cofi-e_-collision_OD--5186`. The setup and theorem assumptions are stated in Sections 2 and 5; protocols, hashes, and regeneration commands are listed in Appendix I.

REFERENCES

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. arXiv preprint arXiv:1811.00075, 2018. URL https://arxiv.org/abs/1811.00075.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29362–29373, 2022.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

Giacomo Boracchi, Diego Carrera, Cristiano Cervellera, and Danilo Macciò. QuantTree: Histograms for change detection in multivariate data streams. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 639–648, 2018. URL https://proceedings.mlr.press/v80/boracchi18a.html.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

Giovanni Cherubin, Adrian Baldwin, and Jonathan Griffin. Exchangeability martingales for selecting features in anomaly detection. In *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pp. 157–170, 2018.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2552–2573, 2022. URL https://proceedings.mlr.press/v151/choi22a.html.

Kacper P. Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1981–1989, 2015.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 11(4):290–294, 1936.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019. doi: 10.1109/JAS.2019.1911747.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Charalambos Eliades and Harris Papadopoulos. A conformal martingales ensemble approach for addressing concept drift. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 328–346, 2023.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

Parikshit Gopalan, Omer Reingold, Vatsal Sharan, and Udi Wieder. Multicalibrated partitions for importance weights. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 408–435, 2022. URL https://proceedings.mlr.press/v167/gopalan22a.html.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16295–16336, 2024. URL https://proceedings.mlr.press/v235/greenblatt24a.html.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Martin Jaggi. Revisiting Frank–Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani. Sequential covariate shift detection using classifier two-sample tests. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9845–9880, 2022.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. iDECODe: In-distribution equivariance for conformal out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7104–7114, 2022. doi: 10.1609/aaai.v36i7.20670.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24749–24769, 2024.

Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981. doi: 10.1109/TIT.1981.1056331.

Jonas M. Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf. AutoML two-sample test. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15929–15941, 2022.

Alix Lhéritier and Frédéric Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018. doi: 10.1109/TIT.2018.2800658.

Myrto Limnios, Nathan Noiry, and Stephan Clémençon. Learning to rank anomalies: Scalar performance criteria and maximization of two-sample rank statistics. In *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 154 of *Proceedings of Machine Learning Research*, pp. 63–75, 2021.

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

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6316–6326, 2020. URL https://proceedings.mlr.press/v119/liu20m.html.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Natalia Martinez Gil, Dhaval Patel, Chandra Reddy, Giri Ganapavarapu, Roman Vaculin, and Jayant Kalagnanam. Identifying homogeneous and interpretable groups for conformal prediction. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2471–2485, 2024.

Natalia Martinez Gil, Fearghal O'Donncha, Wesley M. Gifford, Nianjun Zhou, Dhaval C. Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7uFbs68MSI.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 304–313, 2016. URL https://proceedings.mlr.press/v48/menon16.html.

Nour Moustafa and Jill Slay. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6, 2015.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 577–585, 2016.

Teodora Pandeva, Tim Bakker, Christian A. Naesseth, and Patrick Forré. E-valuating classifier two-sample tests. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=dwFRov8xhr.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49459–49492, 2025.

Aleksandr Podkopaev and Aaditya Ramdas. Sequential predictive two-sample and independence testing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53275–53307, 2023.

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, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/31421b112e5f7faf4fc577b74e45dab2-Abstract-Conference.html.

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, pp. 1394–1406, 2019.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Nabeel Seedat, Alan Jeffares, Fergus Imrie, and Mihaela van der Schaar. Improving adaptive conformal prediction using self-supervised learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10160–10177, 2023. URL https://proceedings.mlr.press/v206/seedat23a.html.

Shalev Shaer, Yarin Bar, Drew Prinster, and Yaniv Romano. Testing for distribution shifts with conditional conformal test martingales. arXiv preprint arXiv:2602.13848, 2026. URL https://arxiv.org/abs/2602.13848.

Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2024. doi: 10.1109/TIT.2023.3305867. URL https://arxiv.org/abs/2112.09162. The linked arXiv version incorporates the 2025 corrections.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18168–18180, 2022.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2(2): 229–260, 2024. doi: 10.51387/23-NEJSDS51.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527, 2025. URL https://proceedings.mlr.press/v267/thurin25a.html.

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, pp. 2526–2536, 2019.

Jean Ville. *Étude critique de la notion de collectif*. Number 3 in Monographies des probabilités. Gauthier-Villars, Paris, 1939.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, volume 60 of *Proceedings of Machine Learning Research*, pp. 132–153, 2017. URL https://proceedings.mlr.press/v60/volkhonskiy17a.html.

Vladimir Vovk. Conformal testing in a binary model situation. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 131–150, 2021.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 768–775, 2003.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 191–210, 2021. URL `https://proceedings.mlr.press/v152/vovk21b.html`.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. Conformal testing: Binary case with Markov alternatives. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pp. 207–218, 2022.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97 (R.1), Hewlett-Packard Laboratories, 2003.

Xilie Xu, Jingfeng Zhang, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli. Adversarial attack and defense for non-parametric two-sample tests. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24743–24769, 2022. URL `https://proceedings.mlr.press/v162/xu22m.html`.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, volume 24, pp. 594–602, 2011.

Zhijian Zhou, Jie Ni, Jia-He Yao, and Wei Gao. On the exploration of local significant differences for two-sample test. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5262–5300, 2023.

The appendix provides notation (Appendix A), implementation details (Appendix B), literature and deployment scope (Appendices C–D), transfer guarantees and proofs (Appendices E–F), and the experimental protocols and complete comparisons (Appendix G). Appendix H summarizes earlier development evidence; Appendix I lists verification commands.

A NOTATION

Table 2 collects the symbols used in the paper. Related and potentially confusable symbols are grouped together. Superscripts such as Q^{Π} denote the law induced by a representation, not exponentiation; hats and underlines denote empirical quantities and certified lower bounds. Symbols local to one proof or experiment are defined where introduced.

Table 2: Notation at a glance.

Group	Symbol	Meaning
Laws	$P,\ Q,\ Q_{1:J}$	null law, one alternative, and the declared J-member portfolio
Observation	$Z \in \mathcal{Z}$	observation and its observation space
Representation	$\Pi = \{C_1, \ldots, C_L\}$	score-group partition with L groups C_ℓ
Representation	$s_\Pi,\ \Pi' \succeq \Pi$	categorical score and a nested refinement
Group laws	$p_\ell,\ q_\ell$	normal and target masses $P(C_\ell)$ and $Q(C_\ell)$
Induced laws	$P^\Pi,\ Q^\Pi$	normal and target categorical score-group laws
Rank	$\mathcal{I}_\ell,\ U_{\mathrm{tie}},\ R_\Pi,\ U$	score-group interval, tie randomizer, population rank, and uniform law
Runtime	$B,\ R_{t,B},\ b_t,\ e_t,\ E_t$	bundle size, finite-bundle rank, evidence function, one-step e-value, and cumulative evidence
Alarm	$\tau_\alpha,\ \alpha$	stopping time and anytime false-alarm level
Information	$D_{\mathrm{KL}},\ \chi^2,\ \mathrm{TV}$	KL divergence, Pearson divergence, and total variation
Density	$r = dQ/dP,\ \mathcal{G}_\Pi = \sigma(\Pi)$	likelihood ratio and partition-generated sigma-field
Decomposition	$\mathrm{Vis}_\Pi(Q)$	Pearson information represented by the score groups
Decomposition	$\mathrm{Hid}_\Pi(Q)$	Pearson information still hidden within score groups
Refinement	$A \subset C,\ G_\Pi(A;Q)$	candidate division inside one score group and its exact information gain
Refinement	$\gamma,\ S_\Pi$	weak-witness edge and total portfolio collision
Dictionary	$\mathcal{A}_\Pi,\ \varepsilon_{\mathcal{A}}$	admissible split class and weak-edge approximation error
Audit	$\underline{G},\ \delta,\ N_{\mathrm{prop}}$	certified gain lower bound, audit error level, and number of proposals
Certified model	$m,\ \Pi_m$	accepted-prefix length and final certified partition
Experiment	K	target cell count or cell cap in the matched comparison
Robust fitting	$\mathcal{M},\ V_\Pi(\mathcal{M}),\ q^\star$	mixture set of score-group laws, its worst-case Pearson value, and least-favorable law
Transfer	$\beta = \min_\ell p_\ell,\ \varepsilon$	minimum normal score-group mass and mixture-coverage radius
Transfer	$b_\eta,\ g_{\mathcal{M}},\ M_{\eta,\beta}$	identity-mixed evidence function, worst-case growth, and log-payoff bound

B EXTENDED THEORY AND IMPLEMENTATION

Technical convention. We assume that $\mathcal{Z}$ is a standard Borel space. This routine measurability condition covers the continuous and discrete spaces used in the experiments and ensures that the conditional distributions in the theory exist.

B.1 WEAK-WITNESS REFINEMENT RATE

Theorem 2 (Weak-witness refinement rate). *Assume $dQ_j/dP \in L^2(P)$ for every j, so $S_{\Pi_0} = \sum_j \mathrm{Hid}_{\Pi_0}(Q_j) < \infty$. Suppose that, whenever $\mathrm{Hid}_\Pi(Q_j) > 0$, the dictionary satisfies*

$$\sup_{A \in \mathcal{A}_\Pi} G_\Pi(A; Q_j) \geq \gamma \, \mathrm{Hid}_\Pi(Q_j)$$

for a fixed $\gamma \in (0, 1]$. Put $S_\Pi = \sum_{j=1}^J \mathrm{Hid}_\Pi(Q_j)$. Then population CoFi-E obeys

$$S_{\Pi_k} \leq (1 - \gamma/J)^k S_{\Pi_0}, \qquad \max_j \mathrm{Hid}_{\Pi_k}(Q_j) \leq S_{\Pi_k}. \tag{11}$$

If the weak edge has additive error $\varepsilon_{\mathcal{A}}$, the recursion has limiting floor $J\varepsilon_{\mathcal{A}}/\gamma$.

The weak-witness condition says that the best available division exposes at least fraction γ of any remaining hidden information (cf. weak-learning edges in boosting; Freund & Schapire, 1997). Thus total hidden information decreases geometrically. The approximation error $\varepsilon_{\mathcal{A}}$ sets the eventual floor.

B.2 FINITE-ALPHABET COMPLETION AND SCORE-GROUP COMPLEXITY

Corollary 2 (Finite-alphabet weak edge). *Suppose $\mathcal{Z}$ is finite, the alternative portfolio is finite, and the dictionary contains every nontrivial binary subset of every current score group. Then a constant $\gamma > 0$ satisfying Theorem 2 exists. Starting from one score group, population CoFi-E reaches $\mathrm{Hid}_\Pi(Q_j) = 0$ for every portfolio member after at most $|\mathcal{Z}| - 1$ refinements.*

Restrict the finite alphabet to the support of P. Whenever $\mathrm{Hid}_\Pi(Q_j) > 0$, the conditional laws differ inside some nonsingleton score group, so some binary subset has positive gain by Theorem 1. There are only finitely many partitions, portfolio members, and binary splits. The minimum positive ratio between the best split gain and $\mathrm{Hid}_\Pi(Q_j)$ is therefore strictly positive. Each accepted division adds one score group, and the all-singleton partition has zero collision. The weak edge is thus automatic in this exhaustive finite setting, though its value may be small.

Proposition 2 (Tight score-group complexity). *Let P be uniform on M symbols and include every pairwise mass-transfer alternative $Q_{ab} = P + \epsilon(e_a - e_b)$ with $0 < \epsilon < 1/M$, where e_a is the unit mass vector at symbol a. Every partition with fewer than M score groups has maximum KL deficiency at least*

$$d_M(\epsilon) = (M^{-1} + \epsilon)\log(1 + M\epsilon) + (M^{-1} - \epsilon)\log(1 - M\epsilon) > 0.$$

Starting from one score group, at least $M - 1$ binary refinements are therefore required to eliminate worst-case deficiency. The completion count is minimax tight.

Appendix F gives the proof.

B.3 ROBUST HULL SELECTION

For score-group law p and training laws $q_1, \ldots, q_J$, define

$$V_\Pi(\mathcal{M}) = \inf_{q \in \mathcal{M}} \left(\sum_{\ell=1}^L q_\ell^2 / p_\ell - 1 \right), \quad \mathcal{M} = \mathrm{co}\{q_1, \ldots, q_J\}. \tag{12}$$

$\mathcal{M}$ contains every mixture of training score-group laws; V_Π is their smallest Pearson separation from the null. Our robust variant ranks feasible splits by a worst-leave-one-out mean-vertex surrogate and solves this simplex quadratic program only for the best 16 candidates. It selects lexicographically: first by minimum inner leave-one-family-out hull gain, then by full-portfolio hull gain, then by mean vertex gain. The audit certificate stays an admissibility gate only. A Frank-Wolfe initialization (Jaggi, 2013) is polished by a constrained simplex solver, and an independent verifier recomputes each optimum.

B.4 FINITE-BUNDLE TRANSFER STATEMENT

Proposition 3 (Finite-bundle transfer). *Fix B, a partition, and a rank evidence function b with $\int_0^1 b \leq 1$ and $0 < m_b \leq b \leq M_b < \infty$. Under an alternative, let the monitored query have*

score-group law q and let the B references be iid with normal score-group law p, independent of the query. Put $h_{B,\ell} = \mathbb{E}_P[\log b(R_{t,B}) \mid Z_t \in C_\ell]$ and $H_b = \max\{|\log m_b|, |\log M_b|\}$. Under the hull-radius premise of Theorem 3,

$$\mathbb{E}_{q_\star} \log b(R_{t,B}) \geq \inf_{q \in \mathcal{M}} q^\top h_B - H_b \varepsilon. \tag{13}$$

Moreover, $D_{\mathrm{KL}}(Q^{R_{t,B}} \| U) \leq D_{\mathrm{KL}}(q \| p)$.

The finite-bundle baseline in (13) is not the population quantity $g_{\mathcal{M}}$ from Theorem 3; the proposition gives a sufficient growth bound, not finite-B monotonicity under refinement or a detection-delay theorem. Appendix F gives the exact multinomial-kernel proof.

B.5 Adaptive evidence rule

Let $K_{B,\hat{p},\ell}(u)$ be the finite-reference rank density conditional on score group ℓ, computed from a frozen positive estimate $\hat{p}$ of the normal score-group probabilities. Each kernel integrates to one. If $n_{t-1,\ell}$ is the number of earlier query observations in score group ℓ, the KT estimate and its evidence function are

$$\widehat{q}_{t,\ell} = \frac{n_{t-1,\ell} + 1/2}{t - 1 + L/2}, \qquad b_t(u) = (1 - \eta) + \eta \sum_{\ell=1}^{L} \widehat{q}_{t,\ell} K_{B,\hat{p},\ell}(u).$$

The current score-group count is added only after its rank is scored. Thus b_t is predictable and integrates to one for every history, even when $\hat{p}$ is inaccurate. Counts are updated from score groups while factors are evaluated on ranks, so the standard KT coding regret bound does not apply directly.

B.6 Witness, selection, and audit roles

1. **Witness data** fit a global cross-fitted logistic-halfspace dictionary; frozen halfspaces are then restricted to current score groups, using classifiers only as proposal generators (Lopez-Paz & Oquab, 2017; Menon & Ong, 2016). Axis thresholds are kept as fallbacks. This stage may generate many candidates.
2. **Selection data** choose a complete greedy path using empirical versions of (7). No audit observation is touched.
3. **Audit data** evaluate only the fixed proposed path. The final detector is the longest prefix in which every split has a simultaneous positive lower certificate and the declared minimum null support.

B.7 Proof of Corollary 1

The first statement is Lemma 2. For an audited target, a positive certificate implies $q_C > 0$ and $P(A \mid C) \neq Q(A \mid C)$, so the Bernoulli conditional-KL term contributed by that split in (10) is strictly positive; all other refinement increments are nonnegative. This is a representation guarantee, not a claim that a newly fitted finite-B evidence function must have nondecreasing growth.

C Extended Related-Work Comparison

Conformal sequential evidence and calibration. Exchangeability martingales and conformal test martingales provide anytime-valid evidence from ranks (Vovk et al., 2003; Volkhonskiy et al., 2017; Vovk, 2021; Vovk et al., 2021; 2022), and e-detectors extend betting-based change detection beyond ranks (Shin et al., 2024). Ensembles and weighted martingales alter the evidence layer (Eliades & Papadopoulos, 2023; Prinster et al., 2025), while conditional and adaptive conformal methods modify calibration, thresholds, weights, or groupwise validity under distribution shift (Tibshirani et al., 2019; Shaer et al., 2026; Gibbs & Candès, 2021; Bastani et al., 2022; Martinez Gil et al., 2026). These methods can substantially improve deployment, but none can recover information that is absent from the score-rank interface it receives. CoFi-E is complementary: it changes the representation and then hands a frozen score to a runtime mechanism that must establish its own conditional e-value premise.

Learning scores and representations. Feature-selection exchangeability martingales, learned conformity scores, self-supervised anomaly scores, learned groups, and score rectification improve ranking, prediction efficiency, or conditional validity (Cherubin et al., 2018; Limnios et al., 2021; Stutz et al., 2022; Kiyani et al., 2024; Kaur et al., 2022; Martinez Gil et al., 2024; Plassier et al., 2025; Seedat et al., 2023; Thurin et al., 2025). One-class methods instead learn deviation from clean support without a declared alternative portfolio (Liu et al., 2008; Schölkopf et al., 2001; Ruff et al., 2018). CoFi-E optimizes a different quantity: alternative information hidden *within the fibers of the current score*. Nested refinement makes already exposed partition information monotone, while the Pearson residual supplies a local observable objective.

Two-sample tests, density ratios, and localized partitions. Classifier and kernel two-sample tests learn global observation-space discrimination (Lopez-Paz & Oquab, 2017; Chwialkowski et al., 2015; Liu et al., 2020; Kübler et al., 2022; Shekhar et al., 2022); proper-loss and density-ratio viewpoints connect such classification to likelihood-ratio estimation (Menon & Ong, 2016; Yamada et al., 2011; Choi et al., 2022). Sequential two-sample and evidence rule procedures pursue direct monitoring power (Lhéritier & Cazals, 2018; Shekhar & Ramdas, 2024; Pandeva et al., 2024; Jang et al., 2022; Podkopaev & Ramdas, 2023). QuantTree constructs null-controlled histograms (Boracchi et al., 2018), and local-significance trees partition a representation to identify where samples differ (Zhou et al., 2023). CoFi-E may use classifiers to propose cuts, but their fitted probabilities are never trusted as certificates: independent counts certify the local split gain, and the output is consumed as a nested conformal rank representation rather than as a global classifier statistic. Dataset-shift detectors likewise test a reduced representation (Rabanser et al., 2019), and two-sample tests can be attacked directly (Xu et al., 2022); RFCA formalizes the case in which a change is invisible to the deployed conformal score itself.

Closest local-refinement precedent. The closest algorithmic overlap is multicalibrated partition refinement (Hebert-Johnson et al., 2018; Gopalan et al., 2022). Its weak learner searches for conditional discrepancies inside current cells, its analysis quantifies divergence gained by splitting, and the full construction can merge cells to control size. CoFi-E applies local refinement to a deployed conformal interface: it keeps an inherited partition, shares one nested path across declared alternatives, certifies each retained cut with independent counts, and passes the resulting partition to a valid rank e-process. The merge-blindness proposition and the joint refinement lemma make the interface distinction explicit, while the Pearson residual targets exactly the information erased by the current score.

D Extended Limitations and Deployment Scope

Summary of empirical qualifications. Retention concerns population information, not monotone finite-reference power, and the rate needs a useful witness. Unseen changes require sufficient score-law coverage by the design family (Appendix E); fresh references and audit data add cost. Historical feature projections make some audits conditional development diagnostics, and four dataset clusters limit bootstrap inference. Collision claims are interface-specific. The paper does not establish Pearson-specific search or repeated-repair superiority, or physical feasibility of intrusion transports. LLM agents motivate the problem but are not evaluated; history-dependent deployments require additional conditional assumptions.

Language-model agents as a motivating application. Model-based monitoring and conformal output guarantees offer different forms of oversight for language models (Greenblatt et al., 2024; Quach et al., 2024). Our motivating example concerns a sequential detector that receives a fixed score of an action or interaction record. A score-law collision would prevent that interface from distinguishing two behavioral distributions; it need not defeat a monitor that reads the full record. Applying CoFi-E would require features that expose the relevant difference, examples of the target change, independent audit data, and a runtime protocol satisfying conditional exchangeability. Dependence within agent trajectories and adaptation to monitoring feedback are not covered by an iid deployment argument. We provide no LLM-agent experiment or guarantee about harmful outcomes; the application motivates the representation-repair question rather than supplying empirical evidence for it.

Power is conditional on coverage. CoFi-E is not a universal anomaly detector. Its rate assumes a weak witness edge for declared alternatives. Its unseen-family theorem requires a score-group-law coverage radius. The multi-domain result shows these assumptions cannot be replaced by optimism about generic transfer.

Static design alternatives. The theory treats a finite portfolio or a covered convex score-group-law family. A history-adaptive adversary would require conditional-kernel coverage, not marginal samples. Conditional RFCA here uses a declared finite context and makes no claim about the full filtration.

Declared interface. The collision definition can be extended to a detector that also observes a finite context H by preserving the joint law of (H, Π), or to several scores by preserving their complete joint score law. These extensions state precisely which interface is blind. A collision for a finite context need not defeat a detector with a richer history, and a statistical transport is not automatically a physically realizable attack.

Reference cost. The exact deployment statement uses fresh conditionally exchangeable bundles. A fixed calibration set needs its own calibration-conditional martingale analysis; treating it as an infinite population is invalid. CoFi-E is compatible with a valid fixed-reference method but does not supply one here.

Statistical and computational scaling. Bonferroni-corrected exact intervals are conservative when many weak proposals are audited; step-down or closed-testing procedures (Holm, 1979; Marcus et al., 1976) could reduce this cost. Hard nested partitions can require many score groups (Proposition 2). Halfspace generation and per-candidate convex programs are expensive in high dimensions. Neural or kernel witnesses remain usable only if generated without audit leakage and frozen before runtime calibration.

Empirical scope. The broad experiment is a design-family transfer study, not a sequential benchmark, and AUROC does not evaluate anytime false-alarm control. The controlled sequential study is exact but low-dimensional. We report both because merging their conclusions would overstate the evidence. The transfer audit resamples only 13 dataset clusters, and its $\hat{\varepsilon}$ values describe reserved empirical score-group laws, not certified population radii.

E A Quantitative Family-Transfer Boundary

The family-coverage assumption cannot be dropped. An unseen alternative can have exactly the normal score-group law even when every training family is well separated. The next theorem gives sufficient quantitative conditions for transfer; the bounds provide a conservative coverage penalty.

Theorem 3 (Convex-hull transfer)**.** *Let $\beta = \min_\ell p_\ell > 0$, let $\mathcal{M}$ be the training mixture set, and suppose an unseen score-group law $q_\star$ satisfies $\inf_{\bar{q}\in\mathcal{M}} \|q_\star - \bar{q}\|_1 \le \varepsilon$. Then*

$$\chi^2(q_\star\|p) \ge V_\Pi(\mathcal{M}) - 2\varepsilon/\beta, \tag{14}$$

$$D_{\mathrm{KL}}(q_\star\|p) \ge \tfrac{1}{2}\left[\sqrt{\beta V_\Pi(\mathcal{M})} - \varepsilon\right]_+^2. \tag{15}$$

Fix an evidence-function law $q^{\mathrm{b}} \in \mathcal{M}$ and $0 < \eta < 1$. For the identity-mixed evidence function $b_\eta = (1-\eta) + \eta q^{\mathrm{b}}/p$, let

$$g_{\mathcal{M}} = \inf_{q\in\mathcal{M}} \mathbb{E}_q \log b_\eta, \qquad M_{\eta,\beta} = \max\bigl\{|\log(1-\eta)|,\, \log(1-\eta+\eta/\beta)\bigr\}.$$

Then

$$\mathbb{E}_{q_\star} \log b_\eta \ge g_{\mathcal{M}} - M_{\eta,\beta}\varepsilon. \tag{16}$$

Here β is the smallest normal score-group mass; ε bounds the sum of absolute probability differences (ℓ_1 distance) to a training mixture. $[x]_+ = \max\{x, 0\}$. The ratio q^{b}/p is computed score group by score group. The fixed law q^{b} supplies the bet, $g_{\mathcal{M}}$ is its worst training-mixture log growth, and

$M_{\eta,\beta}$ bounds its absolute log payoff. Thus the last bound subtracts a coverage penalty from training growth. Proofs are in Appendix F.

Local certificates establish separation for a targeted training alternative. Transfer additionally requires a separated training hull and an unseen law close enough to it. A positive local gain alone establishes neither condition.

F Full Proofs

F.1 Proof of Proposition 1 and Lemma 1

Condition on the frozen design and the past. The iid-query and fresh-reference assumptions, together with $Q^{s_j} = P^{s_j}$, make each current score bundle exchangeable under Q. Its randomized rank is therefore conditionally uniform. Thus any normalized predictable evidence function has conditional expectation at most one under Q: writing $M_j(t)$ for component j's cumulative wealth and $\mathcal{F}_{t-1}$ for the filtration generated by both streams' history through time $t-1$, $\mathbb{E}_Q[M_j(t) \mid \mathcal{F}_{t-1}] \le M_j(t-1)$ for every t, so each M_j is individually a nonnegative supermartingale started at one - an e-process, not merely a fixed-time e-value. A fixed-weight convex combination $M(t) = \sum_j w_j M_j(t)$ ($w_j \ge 0$, $\sum_j w_j = 1$) is adapted to $\mathcal{F}_t$, and by linearity of conditional expectation $\mathbb{E}_Q[M(t) \mid \mathcal{F}_{t-1}] = \sum_j w_j \mathbb{E}_Q[M_j(t) \mid \mathcal{F}_{t-1}] \le \sum_j w_j M_j(t-1) = M(t-1)$, so M is itself a nonnegative supermartingale started at one and Ville's inequality applies to its whole path, giving the alarm bound in Proposition 1 at any stopping time. This is the argument for fixed-weight process mixtures specifically - the only merging rule used anywhere in this paper's experiments - and is what Proposition 1 now states; a merging rule shown valid only as an e-value at one fixed time does not automatically inherit the pathwise supermartingale property this argument uses, and no such broader claim is made.

The joint score (Π_0, Π_1) is a measurable function of any partition Π refining both component partitions. Data processing (Cover & Thomas, 2006) therefore gives

$$D_{\mathrm{KL}}(Q^{\Pi} \| P^{\Pi}) \ge D_{\mathrm{KL}}\left(Q^{(\Pi_0,\Pi_1)} \middle\| P^{(\Pi_0,\Pi_1)}\right).$$

Applying the partition-rank identity proves Lemma 1. For a fully supported $L_0 \times L_1$ joint table, the row and column constraints have rank $L_0 + L_1 - 1$; after the total-mass constraint, the remaining dimension is $(L_0 - 1)(L_1 - 1)$. A joint refinement records the table itself and removes these marginally invisible directions.

F.2 Proof of Lemma 2

Let $\mathcal{I}_\ell$ be the rank interval assigned to score group C_ℓ. Its Lebesgue length is p_ℓ. Conditional on $Z \in C_\ell$, randomized tie smoothing is uniform on $\mathcal{I}_\ell$. Under Q, the interval carries total mass q_ℓ; hence its density relative to Lebesgue measure is q_ℓ / p_ℓ. Therefore

$$D_{\mathrm{KL}}(Q^{R_\Pi} \| U) = \sum_{\ell=1}^{L} \int_{\mathcal{I}_\ell} \frac{q_\ell}{p_\ell} \log \frac{q_\ell}{p_\ell}\, du = \sum_{\ell=1}^{L} q_\ell \log(q_\ell / p_\ell).$$

This is $D_{\mathrm{KL}}(Q^{\Pi} \| P^{\Pi})$. If Π' refines Π, apply the KL chain rule to the coarse score group and its child group to obtain (10); every conditional KL term is nonnegative.

For runtime validity, condition on the pre-bundle filtration and all frozen design data. Conditional joint exchangeability and the query-inclusive tie count in (1) give $R_{t,B} \sim \mathrm{Unif}(0,1)$ conditionally. Thus

$$\mathbb{E}_P[e_t \mid \mathcal{F}_{t-1}] = \int_0^1 b_t(u)\, du \le 1.$$

The product E_t is a nonnegative supermartingale starting at one, and Ville's inequality proves (2). □

F.3 Proof of Theorem 1

Write $\mathcal{G}_\Pi = \sigma(\Pi)$ for the information recorded by the score groups. The conditional expectation $r_\Pi = \mathbb{E}_P[r \mid \mathcal{G}_\Pi]$ equals q_ℓ/p_ℓ on C_ℓ. Consequently,

$$\mathbb{E}_P[(r_\Pi - 1)^2] = \sum_\ell p_\ell (q_\ell/p_\ell - 1)^2 = \sum_\ell q_\ell^2/p_\ell - 1 = \mathrm{Vis}_\Pi(Q).$$

The two terms in $r - 1 = (r - r_\Pi) + (r_\Pi - 1)$ are orthogonal in $L^2(P)$. Pythagoras gives the projection form of hidden information:

$$\mathrm{Hid}_\Pi(Q) = \mathbb{E}_P[(r - r_\Pi)^2], \qquad \chi^2(Q\|P) = \mathrm{Hid}_\Pi(Q) + \mathrm{Vis}_\Pi(Q).$$

If $\Pi' \succeq \Pi$, then $L^2(\mathcal{G}_\Pi) \subseteq L^2(\mathcal{G}_{\Pi'})$. Applying Pythagoras to the two projections yields (6).

Only the parent term changes after splitting C into A and $C \setminus A$. The increase in visible information is

$$\frac{q_A^2}{p_A} + \frac{(q_C - q_A)^2}{p_C - p_A} - \frac{q_C^2}{p_C} = \frac{(p_C q_A - q_C p_A)^2}{p_C p_A (p_C - p_A)}.$$

This proves (7).

$\mathrm{Hid}_\Pi(Q) = 0$ iff $r = r_\Pi$ P-almost surely. On every Q-positive score group, this means $dQ/dP = q_\ell/p_\ell$, equivalent to $Q(\cdot \mid C_\ell) = P(\cdot \mid C_\ell)$. Conversely conditional-law equality makes r score-group measurable. If every set in a measure-determining class has zero gain, the numerator of (7) vanishes and the conditional laws agree on that class and hence agree by its measure-determining property. Affine halfspaces determine Borel probability laws on $\mathbb{R}^d$ by Cramér-Wold (Cramér & Wold, 1936). □

F.4 Proof of Theorem 2

For $S_\Pi = \sum_j \mathrm{Hid}_\Pi(Q_j)$, $\max_j \mathrm{Hid}_\Pi(Q_j) \geq S_\Pi/J$. The weak witness condition and exact greedy choice therefore give

$$G_{\Pi_k}(A_k; Q_{j_k}) \geq \gamma S_{\Pi_k}/J.$$

The selected collision falls by exactly this amount, while Theorem 1 implies every unselected collision is nonincreasing. Hence

$$S_{\Pi_{k+1}} \leq (1 - \gamma/J) S_{\Pi_k},$$

and induction proves (11). With additive dictionary error,

$$S_{k+1} \leq (1 - \gamma/J) S_k + \varepsilon_{\mathcal{A}},$$

whose stationary floor is $J\varepsilon_{\mathcal{A}}/\gamma$. □

F.5 Validity of the universal evidence function

At time t, $\widehat{q}_t$ depends only on query score groups before the current arrival and is therefore predictable. Since $\int K_{B,\hat{p},\ell}(u)\, du = 1$ for every score group and every frozen positive estimate $\hat{p}$,

$$\int_0^1 b_t(u)\, du = (1 - \eta) + \eta \sum_\ell \widehat{q}_{t,\ell} = 1.$$

The conditional uniformity of the fresh-bundle rank then gives $\mathbb{E}[b_t(R_{t,B}) \mid \mathcal{F}_{t-1}] = 1$ under the null. The identity mixture keeps every factor at least $1 - \eta > 0$. Because counts are updated from query score groups while log loss is evaluated on randomized ranks, the usual KT coding inequality does not directly establish a regret bound for this process. Null-law estimation can affect its alternative growth without changing the normalization argument. □

F.6 Proof of Proposition 2

A partition with fewer than M score groups contains distinct a, b in one score group. The perturbation $Q_{ab} - P = \epsilon(e_a - e_b)$ transfers mass only inside this score group, so $Q_{ab}^\Pi = P^\Pi$ and its visible rank KL is zero. Direct substitution gives

$$D_{\mathrm{KL}}(Q_{ab}\|P) = d_M(\epsilon) > 0.$$

Thus maximum deficiency is at least $d_M(\epsilon)$. Starting from one score group, a binary refinement adds at most one score group, so at least $M - 1$ updates are necessary. □

F.7 Proof of the audit lower bound

Use $p_A = p_C u$ and $q_A = q_C v$ in (7):

$$G_\Pi(A;Q) = \frac{q_C^2}{p_C}\frac{(v-u)^2}{u(1-u)}.$$

On the simultaneous interval event, $q_C \geq \underline{q}_C$, $p_C \leq \overline{p}_C$, and $|v-u| \geq d(I_u, I_v)$. Since $u(1-u) \leq 1/4$,

$$G_\Pi(A;Q) \geq 4\frac{\underline{q}_C^2}{\overline{p}_C} d(I_u, I_v)^2.$$

Each fixed proposal uses four intervals. If $q_C = 0$, its population gain is zero and the lower certificate is zero on the coverage event; undefined conditional intervals from zero counts are taken as $[0,1]$. Giving each miscoverage probability $\delta/(4N_{\mathrm{prop}})$ and applying a union bound proves simultaneous coverage at least $1-\delta$. □

F.8 Proof of Theorem 3

For probability vectors q, q' and $p_\ell \geq \beta$,

$$\left|\chi^2(q\|p) - \chi^2(q'\|p)\right| \leq \sum_\ell \frac{|q_\ell - q'_\ell|(q_\ell + q'_\ell)}{p_\ell} \leq \frac{2}{\beta}\|q - q'\|_1.$$

Choose a nearby mixture point $\tilde{q} \in \mathcal{M}$ with $\|q_\star - \tilde{q}\|_1 \leq \varepsilon$. Then $\chi^2(\tilde{q}\|p) \geq V_\Pi(\mathcal{M})$ gives (14), and

$$\chi^2(\tilde{q}\|p) = \sum_\ell (\tilde{q}_\ell - p_\ell)^2/p_\ell \leq \|\tilde{q} - p\|_1^2/\beta.$$

Hence

$$\|q_\star - p\|_1 \geq [\sqrt{\beta V_\Pi(\mathcal{M})} - \varepsilon]_+,$$

and Pinsker's inequality (Cover & Thomas, 2006) proves (15).

The evidence function law q^{b} is fixed independently of this nearby point. Because $1-\eta \leq b_\eta \leq 1-\eta+\eta/\beta$, $\|\log b_\eta\|_\infty \leq M_{\eta,\beta}$, and therefore

$$|\mathbb{E}_{q_\star} \log b_\eta - \mathbb{E}_{\tilde{q}} \log b_\eta| \leq M_{\eta,\beta}\|q_\star - \tilde{q}\|_1.$$

Since $\tilde{q} \in \mathcal{M}$ implies $\mathbb{E}_{\tilde{q}} \log b_\eta \geq g_{\mathcal{M}}$, this proves (16). □

F.9 Proof of Proposition 3

Given query score group ℓ, let $N = (N_1, \ldots, N_L) \sim \mathrm{Mult}(B, p)$ be the independent reference counts and $N_{<\ell} = \sum_{i<\ell} N_i$. Conditional on N and the query score group, (1) is uniform on $[N_{<\ell}/(B+1), (N_{<\ell} + N_\ell + 1)/(B+1)]$. Thus

$$h_{B,\ell} = \sum_{n_1+\cdots+n_L=B} \frac{B!\prod_i p_i^{n_i}}{\prod_i n_i!} \frac{B+1}{n_\ell+1} \int_{n_{<\ell}/(B+1)}^{(n_{<\ell}+n_\ell+1)/(B+1)} \log b(u)\, du.$$

This defines a Markov kernel $K_{B,p}$ independent of the query law, so $Q^{R_t,B} = qK_{B,p}$ and $pK_{B,p} = U$. Data processing gives $D_{\mathrm{KL}}(Q^{R_t,B}\|U) \leq D_{\mathrm{KL}}(q\|p)$. Also $\|h_B\|_\infty \leq H_b$, so for any nearby $\tilde{q} \in \mathcal{M}$,

$$q_\star^\top h_B \geq \tilde{q}^\top h_B - H_b\|q_\star - \tilde{q}\|_1 \geq \inf_{q\in\mathcal{M}} q^\top h_B - H_b\varepsilon,$$

which proves (13). □

For example, with $B = 1$, $p = (1/2, 1/2)$, and $q = (1, 0)$, finite-bundle rank KL is 0.1308, whereas population-rank KL is $\log 2 = 0.6931$. Thus the population identity cannot be transferred unchanged to a finite reference bundle.

G Experimental Protocols and Results

The subsections follow the main experiments: controlled repair, real-feature repair, corrected UNSW replay, and the matched and exploratory UNSW studies. Earlier development results are summarized separately in Appendix H.

G.1 Controlled repair and joint evasion

Table 3 evaluates a frozen horizontal score repaired using a hidden vertical change on four points. It reports both population rank information and finite-reference log evidence for the earlier and hidden families. The laws, sampling roles, and exact evaluation are specified below.

Table 3: Controlled detector-repair comparison (five seeds). KL columns are population rank information in nats. g_B is exact expected log evidence per step for the equal mixture of the lifted legacy component and newly fitted hidden-family component, with $B = 63$. Values are seed means; the common cap is four score groups, not equal accepted size.

Method	Score groups	Legacy KL	Hidden KL	Legacy g_B	Hidden g_B
Original detector	2	0.1927	0.0000	+0.1196	−0.0467
CoFi-E certified repair	4	0.1927	0.1927	+0.0972	+0.0600
Unrestricted certified refit	2	0.0000	0.1927	−0.0462	+0.1193

The separate E1 protocol evaluates an ensemble evasion that preserves the legacy and new score marginals while changing their joint law. Its population joint KL is 0.08228 nats and its finite-B product-rank KL is 0.06200 nats. At onset 25, process-level merging detects only 0.0138 of streams, whereas plain CoFi-E detects 0.8414 and CoFi-E+ detects 0.8264. Their null crossing rates are 0.0378, 0.0310, and 0.0344, respectively. CoFi-E+ uses the fixed four-component weights $(0.10, 0.10, 0.40, 0.40)$; plain CoFi-E uses $(0.5, 0.5)$ for its declared-refined and universal-refined components. The fixed-product comparator uses a declared score-group law on the same four-cell map that CoFi-E uses, so this contrast isolates two effects: the joint interface against marginal merging, and adaptive against declared betting. It does not isolate the refinement search or its certificates. Here the certified refinement reproduces the product partition exactly, and a product comparator given the same two-component mixture is plain CoFi-E by construction. The laws, sampling roles, and exact evaluation are specified below.

Four-point design and exact evaluation. The frozen protocol `experiments/existing_score_repair_protocol.json` uses atoms $(-1,-1),(-1,1),(1,-1),(1,1)$ with null law $p = (.25, .25, .25, .25)$, hidden alternative $q_h = (.1, .4, .1, .4)$, and earlier alternative $q_e = (.1, .1, .4, .4)$. The initial score distinguishes $x \le 0$ from $x > 0$. The hidden alternative preserves its two masses exactly; the earlier alternative is visible. Only q_h is used to train the refinements.

Each of five declared seeds draws 3,000 independent observations per law per witness/selection/audit role. Two analytically fixed axis cuts at zero are the shared dictionary. Repair starts from two score groups with at most two new divisions; refitting starts from one score group with at most three. Both use audit level .05, minimum child support 20, final score-group cap four, identity weight .99, and reference size $B = 63$. The fixed dictionary needs no fitted witness direction. All ten proposed repair splits pass the independent count certificate. Refitting accepts one split in each seed. These are equal-cap comparisons; the actual accepted sizes in Table 3 must be retained when interpreting information and growth.

Evaluation enumerates the known atoms and integrates the exact finite-bundle rank kernel against each fitted evidence function; it introduces no evaluation sampling error. The hidden finite-bundle rank KL is 0, .1172, and .1650 for original, repair, and refit. The corresponding earlier-family finite-rank KL is .1650, .1650, and 0. Thus preservation of population information need not maximize finite-reference information or fitted-evidence function growth. The result `results/cofie_existing_score_repair.json` contains all 15 records; `verify/verify_existing_score_repair.py` independently checks the protocol and source hashes, certificates, partitions, rank normalization, information, growth, and summaries.

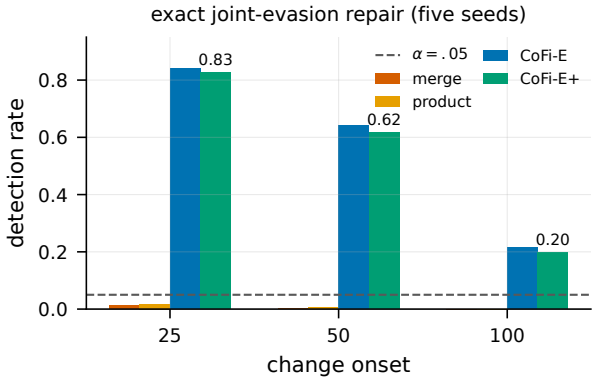


Figure 3: Exact four-point joint-evasion detection across five seeds. The marginal merge stays near its null alarm level, while nested refinement and its universal component retain power as the change arrives later.

Two-family finite-reference repair evaluation. The frozen two-family repair protocol `experiments/detector_repair_comparison_protocol.json` repeats the same four-atom construction with five new seeds, the same 3,000 observations per law and role, the same fixed cuts, four-score-group cap, audit level, and $B = 63$. It carries forward the frozen legacy score-group evidence function, learns only the hidden family from the new role samples, and evaluates their equal-weight mixture. Table 3 gives its exact log-growth values. The analytic normal law normalizes the fitted hidden-family evidence function before kernel evaluation; this is a controlled diagnostic rather than an empirical calibration guarantee. The protocol and independent verifier are released with the five-seed record.

E1 mixture weights and rejected run. The initial frozen E1 run used equal CoFi-E+ weights $(0.25, 0.25, 0.25, 0.25)$ and yielded 0.7956 onset-25 detection (while unchanged plain CoFi-E yielded 0.8414), so it was rejected by Gate A; the promoted rerun changed only the four fixed weights to $(0.10, 0.10, 0.40, 0.40)$, and the rejected protocol and result are preserved with SHA-256 hashes in `docs/rejected-e1.sha256`. That rejected record was regenerated by rerunning the frozen equal-weight protocol, whose fixed seeds and stream design make the run deterministic; it is not a byte copy of the original execution.

G.2 Real-feature repair and intrusion diagnostics

E2b repairs a frozen Isolation Forest (Liu et al., 2008) interface on thirteen feature artifacts. The natural track in Table 1 covers all seven declared dataset–family pairs: one each on MVTec Carpet, UCR Crop, and UEA FaceDetection, and four on UCR ElectricDevices. Five seeds give 35 records, weighted equally; thus ElectricDevices contributes four of the seven tasks. All seven methods and all three frozen onsets are included. Each condition uses 200 streams; evaluation rows are reserved before fitting. Target detection, earlier-family retention, and null alarm rates in the main table all use these same 35 records. HGB denotes histogram gradient boosting. Each family and the null pool are split once per seed into disjoint design and evaluation halves, so every number below is measured on rows that no fitting role saw; five records are excluded because ECG5000 reserves only five legacy rows. Table 4 reports the endpoint. On the natural blind-spot tracks CoFi-E+ reaches onset-25 detection 0.979 with legacy retention 0.974, against 0.977 and 0.957 for the merged pair that its mixture contains; the replacement baselines detect comparably but retain little, 0.253 for the logistic refit. Plain CoFi-E, which carries no legacy component, reaches 0.855. The small onset-25 differences between CoFi-E+ and merging do not establish statistical superiority. The comparison also changes the evidence mixture, so it does not isolate an advantage due to partition search alone.

The collision track perturbs the two interfaces the merge baseline actually deploys, and admits a cycle only when all four corner cells hold at least five rows, share one child label, and straddle the parent cut, so the target table is realizable by reweighting real rows. Under that rule a cycle exists in $3/60$ records: of $10{,}038$ candidate cycles, 76.3% fail corner support, 20.7% fail straddling and

Table 4: E2b real-feature repair with evaluation rows reserved from every fitting role (100 records over thirteen artifacts and five seeds; five excluded where the reserve floor cannot be met). Values are mean onset-25 detection rates; retention is detection of the earlier legacy family. The collision track perturbs the two deployed interfaces and is constructible in $3/60$ records.

Track / metric	IF	Logit	HGB	Merge	Product	CoFi-E	CoFi-E+
Natural detection	.330	.978	.838	.977	.845	.855	**.979**
Natural legacy retention	.936	.253	.360	.957	.708	.971	**.974**
Collision detection (all 60)	.003	.000	.001	.002	.015	.015	.013
Collision detection (3 realized)	.002	.000	.000	.005	.027	.033	.028

2.8% fail purity, with a median best corner count of four against the required five. The three realized attacks are detected in 0.033 of streams by CoFi-E and 0.005 by merging, both below $\alpha = 0.05$, and no dataset exceeds the nominal level. The merge-blindness separation therefore remains a controlled result: the construction succeeds in only three records. The superseded E2 run reported $9/65$ under a weaker construction that preserved a within-parent child indicator rather than the deployed logistic interface, and whose targets could place mass on cells holding no observations.

Intrusion-domain repair. The corrected UNSW-NB15 replay (E3c) supersedes E3b; its corrections are specified in Appendix G.3. Certification remains $102/160$ because the audit uses training-role counts. At onset 25, the adaptive refined component detects 0.756, the declared refined component 0.133, and marginal merging 0.424. At onset 50, adaptive refinement detects 0.710 against merging's 0.106. The refined interface with declared evidence therefore underperforms merging; this experiment alone does not isolate a search advantage. That question motivates Section 6.3.

G.3 UNSW-NB15 REPAIR: CORRECTED ONLINE REPLAY (E3C)

E3/E3b's shared evaluator built the adaptive component's query-score-group input as `ppost[...]` for the entire horizon during hidden-family evaluation and `pclean[...]` for the entire horizon during retention evaluation, selecting by evaluation purpose rather than by arrival time. The true rank stream already routes correctly, via `cofie.sequential._query_labels`: clean before the onset, the evaluated pool from the onset on. E3/E3b's evaluator also estimated each method's normal score-group law from the evaluation-time array rather than freezing it at fit time, and fit preprocessing medians and scales on the full training CSV rather than benign rows only. `cofie_repair/docs/e3b_replay_audit.md` documents the source review and a minimal reproduction; `verify/probe_e3_replay_inputs.py` reproduces the routing error without loading any data.

E3c (`experiments/run_e3c_unsw_nb15_repair.py`) corrects all three: the adaptive component's query label is routed with the same onset-aware formula the rank stream already uses; every null law, target law, and cell alphabet is frozen once from training-role selection data before any testing-CSV row is read; and preprocessing statistics are fit on benign training rows only. E3/E3b's protocol, runner, and result are unchanged and remain historical records; the audit rule is otherwise identical to E3b's. Every stored audit record also carries its Clopper-Pearson totals, so `verify/verify_e3c_unsw_nb15_repair.py` independently reconstructs each lower gain from raw counts, not only the accept/reject threshold against an already-stored value. Twenty-five behavioral tests (`tests/test_e3c_replay.py`) check the corrected routing against `cofie.sequential._query_labels` across randomized configurations, confirm frozen laws are unaffected by which evaluation array is later supplied, and confirm a post-change pool placed beyond the horizon cannot influence any reported value. The full before/after comparison, including the static-AUROC dissociation that explains the power loss without a validity violation, is in `docs/e3c_correction_log.md`.

G.4 MATCHED COMPARISON AND UNSW FOLLOW-UPS

Shared correction record. An independent review of Section 6.3 and the two sections following it found six defects, corrected before the numbers reported in this paper: UNSW preprocessing had been fit on audit rows in addition to witness rows; the audit's simultaneous error budget had been allocated per task rather than per the protocol's stated per-(dataset, seed) scope; the repeated-repair

Table 5: Matched-budget comparison, all rules with the same adaptive evidence function: relative percentage reduction in restricted delay for CoFi-E versus each comparator (higher is better), with 97.5% paired dataset-cluster bootstrap intervals. Fixed grid and Gini are the preregistered primary contrasts, corrected together within each view. Comparisons with the unrepaired detector and random cuts are secondary analyses with pointwise intervals; their observed effects favor CoFi-E in all four datasets. Audit-gated comparisons share a cell cap, not necessarily an accepted cell count.

Partition	Unrepaired	Random cuts	Fixed grid	Gini search
Proposed, exact K	34.8% $[22.0, 44.4]$	34.6% $[25.0, 41.8]$	18.8% $[2.6, 27.7]$	1.1% $[-1.2, 4.5]$
Audit-gated prefix	26.7% $[13.6, 34.4]$	26.7% $[14.3, 33.5]$	24.6% $[13.6, 34.2]$	1.4% $[-4.1, 7.3]$

audit had reused one clean sample across all three rounds without splitting its error budget; the joint-collision experiment's information quantity was mutual information mislabeled as KL divergence; its null-crossing statistic excluded pre-onset alarms; and four random seeds used Python's non-reproducible built-in `hash()`. All six are fixed in `experiments/run_matched_benchmark.py`, `run_repeated_repair.py`, and `run_unsw_joint_collision.py`; the corrected results in the main text supersede an earlier, uncorrected pass whose headline conclusion for the matched comparison happened not to change, but whose repeated-repair result did.

G.4.1 Matched-study protocol and complete comparison

The four-dataset panel contains 290 tasks at cell budgets $K = 8, 16$ and five seeds. Witness, selection, audit, and evaluation roles are fixed before fitting. UNSW preprocessing uses clean witness rows only. For each (dataset, seed), all family/rule/repeat proposal paths are enumerated before audit, with shared budget prefixes counted once; each of their four count intervals receives $0.05/(4N)$, where N is that scope's proposal count. Historical feature projections are held fixed; because they predate these role splits, their audit results are development diagnostics rather than fresh population coverage statements. The three random-cut repeats are averaged within each task. Evidence functions use selection-only fitted laws and $\eta = 0.99$; query/reference draws and tie uniforms are shared across methods within each matched runtime task.

Table 6 gives all six methods under the adaptive evidence function. The registered primary view is the proposed exact-K partition. The success rule also requires positive reductions on at least two datasets, positive common-feasible comparisons, and positive aggregate reductions under the scheduled-start mixture. The required 10% improvement over both comparators already fails, so no success is claimed; the scheduled-start and $B \in \{15, 255\}$ sensitivity runs were not performed. The audit-gated view is reported separately and is not used to replace the registered primary endpoint. Its accepted cell counts can differ between methods.

Interpretation and registered outcome. The Pearson–Gini reductions are 1.1% $[-1.2, 4.5]$ for the proposed exact-K view and 1.4% $[-4.1, 7.3]$ for the audit-gated view, so neither establishes a Pearson-specific advantage. The registered criterion of at least 10% improvement over both primary comparators with intervals excluding zero is not met. Mean accepted splits are 1.68 for Pearson, 2.27 for Gini, and 0.10 for the grid. The population rate does not apply to these estimated implementations without further assumptions. Scheduled-start and reference-size sensitivities were not run; the audit-gated view does not replace the registered exact-K endpoint.

G.4.2 Three-round repair attempt

The family order is DoS, Exploits, Reconnaissance, with seed 202609301. Clean audit blocks contain 4,667, 4,667, and 4,666 observations; each round uses its own family audit sample and $\delta = 0.05/3$. Only DoS accepts cuts, so the final map has seven score groups. Replacement uses the same legacy four-bin start and searches to sixteen score groups using clean and Reconnaissance rows only. It is not a family-specific refit for DoS or Exploits. All arms use $B = 63$, 1,000 streams, and onsets 25, 50, and 100. The runtime streams in this exploratory comparison use separate seeds for different arms; Table 7 is descriptive, not a paired inference result. KT counts are updated online from each evaluated stream, so a family absent from offline fitting can still be learned by the evidence function after its arrival.

Table 6: Restricted delay for all matched-study rules with the adaptive KT evidence function (lower is better). Each row averages families, seeds, cell budgets, and onsets; the final row weights datasets equally.

Dataset	CoFi-E	Gini	Grid	Quantile	Random	Legacy
Proposed partition, exact K						
ElectricDevices	42.18	44.48	54.77	55.17	76.22	80.31
FordA	29.76	30.21	26.72	26.59	38.66	40.23
FaceDetection	64.01	64.12	78.00	75.40	98.91	99.76
UNSW-NB15	44.60	43.80	62.83	62.16	62.51	56.51
Dataset-balanced mean	45.14	45.65	55.58	54.83	69.07	69.20
Audit-gated prefix						
ElectricDevices	50.60	55.52	80.31	78.69	78.69	80.31
FordA	36.38	36.38	40.23	40.23	40.10	40.23
FaceDetection	71.84	72.07	91.95	93.52	99.54	99.76
UNSW-NB15	43.96	41.76	56.51	56.51	58.25	56.51
Dataset-balanced mean	50.69	51.43	67.25	67.24	69.14	69.20

Table 7: Three-round attempt: detection rate and restricted delay in parentheses. Only the first round modifies the nested partition. Merge and product use unsearched components and do not isolate combination strategy.

Family	Nested (7 cells)	Refit (16)	Merge	Product (32)
DoS	0.988 (18.3)	0.974 (22.3)	0.971 (33.7)	0.971 (33.9)
Exploits	0.984 (22.2)	0.978 (37.1)	0.969 (38.0)	0.964 (43.1)
Reconnaissance	0.982 (46.5)	0.973 (39.3)	0.713 (94.6)	0.580 (102.0)

The merge arm averages four KT processes (legacy plus one per family); the product arm uses their static cross. Each new family component uses one global-median division, while the nested and replacement arms use Pearson search. Thus the comparison confounds combination strategy with component strength. Later rounds' failure to certify may reflect smaller gains, limited audit samples, or both; the experiment does not distinguish these explanations.

G.4.3 Approximate joint collision

One fitted legacy score and a new score trained with the Generic family produce a 4×8 table on 14,000 clean construction rows. Of 168 candidate cycles, 129 meet the minimum four-corner support of 20 rows. The selected cycle is chosen from construction counts, then evaluated with $B = 63$ and 1,000 streams per onset. Its target table has exactly the construction marginals. Finite resampling produces row/column ℓ_1 discrepancies of $0.015/0.020$ and KL 0.230 nats relative to the construction table; target KL is 0.232. Mutual information is a different quantity: 0.309 nats for the null table and 0.365 for the realized attack table.

Evaluation references come from the disjoint clean audit pool, whose marginals need not equal those of the construction pool. The reported row/column discrepancies between these pools are $0.027/0.035$. Thus this experiment does not instantiate the exact premise of Proposition 1 at deployment. Table 8 reports attacked-stream detection and an independently simulated null check. Detection counts first alarms from onset through onset$+149$ among all streams, whereas the null check counts every alarm through time 249. Neither is conditioned on surviving without an earlier alarm. Their numerical difference is not an estimate of attack-induced power. No paired no-attack contrast or uncertainty across independently fitted interfaces is reported; sparse occupancy is a possible explanation for low power, not a diagnosis established by this run.

H Earlier Development Studies

These studies preceded the detector-repair experiments. They provide context for the coverage limitation; they are not additional tests of the matched search claim. Full historical tables, plots,

Table 8: UNSW approximate-collision diagnostic. Detection and null crossing have different event definitions and should not be directly subtracted.

Interface	Post-onset detection	Full-horizon null crossing
Legacy score	0.008	0.040
New score	0.009	0.041
Process mixture	0.010	0.041
Joint interface	0.026	0.040

Table 9: Held-out-family AUROC (605 folds; higher is better). Fiber and Robust are CoFi-E variants (ours); OC: oblique collision; PG: pooled-gain partition; IF: Isolation Forest; OCS: one-class SVM. Mean weights datasets equally. Bold values denote column maxima, not statistical significance.

	MVTec			UCR					UEA					
Method	Bottle	Cable	Carpet	Crop	ECG5000	Electric Devices	FordA	UWave Gesture LibraryAll	Face Detection	Heartbeat	Motor Imagery	Self Regulation SCP1	Spoken Arabic Digits	Mean
Fiber (ours)	.754	.536	.564	.754	.559	.689	.629	.546	.492	.498	.508	.501	.536	.582
Robust (ours)	.804	.587	.621	.619	.573	.603	.634	.610	.499	.500	.510	.499	.436	.577
OC	.794	.605	.627	.877	.693	.619	.739	.592	.495	.475	.470	.496	.628	.624
PG	.817	**.631**	.576	.855	.750	.678	.731	.593	.478	**.551**	.458	**.528**	.634	.637
IF	**.969**	.602	**.771**	.981	.961	.755	.935	.990	.597	.419	**.513**	.523	.787	**.754**
OCS	.959	.569	.737	**.987**	**.985**	**.814**	**.939**	**.992**	**.606**	.348	.468	.425	**.804**	.741

protocols, and verification records remain in the repository. The earlier appendix text is archived in `cofie_repair/docs/archive/appendix-before-consolidation.tex`.

Controlled development checks. Earlier Gaussian and finite-alphabet studies checked information recovery, finite-reference normalization, and attacks that preserve marginal, conditional, or joint score interfaces. The Gaussian comparison used known full-observation KL to select a target, rather than Algorithm 1's count-based criterion. These checks do not establish a search advantage for the current certified implementation; Section 6.3 supplies the matched comparison.

Held-out-family transfer. Across 13 datasets (three MVTec categories, five UCR datasets, and five UEA datasets) and five seeds, the historical study evaluated 605 folds, each excluding one complete design family (Bergmann et al., 2019; Dau et al., 2019; Bagnall et al., 2018). The image features used frozen ResNet-50 features with a clean-design projection; time series used deterministic features with a clean-design projection. Synthetic corruptions and eligible training classes supplied the design alternatives; real-defect and held-out-class confirmation arrays remained outside this analysis. The endpoint was static AUROC, not sequential detection.

Fiber CoFi-E certified $1{,}519/2{,}847$ proposals and at least one division in 90.7% of folds, but achieved only 0.5820 macro AUROC. Robust-hull CoFi-E certified $791/2{,}535$ proposals and achieved 0.5765. Both trailed Isolation Forest (0.7540) and one-class SVM (0.7410) (Liu et al., 2008; Schölkopf et al., 2001). Thus a positive design-family certificate did not establish competitive unseen-family ranking.

Reserved-sample coverage audit. A separately frozen resplit reserved each family's evaluation observations before fitting, selection, certification, and evidence-function estimation. For each held-out empirical score-group law, a linear program computed its ℓ_1 distance to the training mixture set. The same reserved observations were evaluated with their family excluded and represented in other portfolios.

All $1{,}815$ Pearson, KL, and growth inequalities held for the empirical laws; these were implementation checks of the theorem, not statistical evidence for its assumptions. Within datasets, radius and log evidence had Pearson correlation -0.717, while the dataset-level Spearman correlation was 0.045. Representing a family raised mean AUROC from 0.5722 to 0.6137: a difference of 0.0415 with 95% dataset-bootstrap interval $[0.0134, 0.0779]$ across 13 clusters. Mean held-out log evidence was 0.0732 nats, with interval $[-0.0382, 0.2103]$ crossing zero. Zero radii could also arise from uninformative representations. These results support a coverage boundary, not generic anomaly-

detection performance. Empirical radii would require a separate sampling certificate for population coverage (Weissman et al., 2003).

An earlier robust-hull run was rejected when independent optimization found a 1.91×10^{-3} discrepancy; solver polishing reduced the verification disagreement to 1.91×10^{-11}. The rejected artifact and historical protocol/result hashes are retained in the release and archived appendix.

I Reproducibility and Verification

All method code, protocols, tests, result JSON, figures, and independent verification scripts are released with the paper. Run the principal commands from the repository root in the frozen research environment. On the study workspace, prefix each command with `bashcofie_repair/hpc/with_env.sh`; the environment is recorded in `cofie_repair/docs/matched-refinement-execution-freeze.json`.

```
python -m pytest tests -q
python verify/verify_cofie_continuous.py
python verify/verify_fiber_witness_development.py
python verify/verify_robust_hull_development.py
python verify/verify_transfer_validation.py
python verify/verify_existing_score_repair.py
python cofie_repair/verify/verify_p3_rank_kernel.py
python cofie_repair/verify/verify_e1_controlled_repair.py
python cofie_repair/verify/verify_e1_rejected_artifact.py
python cofie_repair/verify/verify_e2_real_data_repair.py
python cofie_repair/verify/verify_e2b_real_data_repair.py
python cofie_repair/verify/verify_e3b_unsw_nb15_repair.py
python cofie_repair/verify/verify_e3c_unsw_nb15_repair.py
python -m pytest cofie_repair/tests -q
python cofie_repair/verify/verify_matched_role_manifest.py
python cofie_repair/verify/verify_matched_partition_evaluator.py
python cofie_repair/paper/figures/make_paper_figures.py
```

The new repair records are checked by the independent E1, E2, and P3 verifiers listed in the working-folder README; they verify protocol hashes, source hashes, certificates, normalization, endpoints, and aggregates. The separate rejected-E1 verifier checks the equal-weight protocol, its byte hash, and the recorded 0.7956 and 0.8414 onset-25 endpoints.

The last command regenerates the empirical figures from frozen result JSONs. Figure 1 combines analytic and schematic panels with a runtime simulation; its laws and simulation settings are defined in the figure script.

The robust protocol SHA-256 is `a92e8ffcb4ad02821898eba1ada669bd15adaf4fd3b7198ce228662ec3d6c290`; the corrected result SHA-256 is `73bdca93e6734a1de4a1ae43dae34794cf0f2f45f6cca29e33eb576ed9b235fd`. The transfer protocol and result hashes are `3676935a9dc1dcb8cfae8404f3f02e62965ffcde352ccad764ee5688d04a3670` and `e0c6f2c2833759b9c511efe2d175c720314598ce9a2d0d74c7e4c1bbf4ecd563`. The four historical result verifiers listed above pass $49/49$ checks (continuous: 13; fiber, robust hull, and transfer: 12 each). The current four-point recovery verifier adds $9/9$ checks, while the independent E1 and E2 repair verifiers are listed with their protocols above. The frozen baseline ledger records $68/68$ implementation tests; the repair-study ledger records $57/57$. These are separate suites, not a combined count. Frozen historical ledgers retain their stage-specific totals.