SAMPLING THE SHELL: FROM EXPONENTIAL TO CONSTANT SAMPLE COMPLEXITY IN NEGATIVE-SAMPLING ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Negative-sampling anomaly detection reduces unsupervised fault detection to binary classification: observed telemetry is labeled normal, synthetic negatives are drawn from a reference measure, a classifier is trained, and Integrated Gradients turns alarms into per-sensor symptoms (Sipple, 2020; Sundararajan et al., 2017; Sipple & Youssef, 2022). The recipe is deployed in intelligent diagnostics, yet no theory has said when, and at what dimension, it can work. We supply one, grounded in high-dimensional probability (van Handel, 2016). *Limits:* we identify that manufactured negatives are ancillary: no negative-generation scheme adds boundary information, so $M \gtrsim \rho/\varepsilon$ positives are necessary at resolution ε and occupancy ρ, and uniform sampling operates only inside a dimension window with closed-form optimum $D^\star$ obeying a $1/e$ occupancy law, budgets linear in $1/\varepsilon$ inside and exponential outside. *Constructions:* we propose shell-concentrated, importance-weighted samplers that restore a dimension-free band-hit rate, an exponential-to-constant reduction in negative sample complexity, with data-seeded Langevin dynamics mixing in dimension-free time and finite-sample certificates. *Algorithms:* we construct a ladder of eight detectors, from the MADI baseline to certified latent pipelines, each motivated by a theorem, including an end-to-end certificate for the load-bearing rung that covers estimated boxes and exact finite-sample draws with no equilibration assumption. Experiments track the theory, with bootstrap intervals stating what each sample size resolves. A matched-budget dimension sweep observes the predicted collapse of uniform sampling outside its window while certified boxes hold ROC-AUC 1.000 through $D = 128$. On synthetic suites, mode shape picks the rung: certified boxes are exact on box modes, frozen-encoder rungs win rotated-Gaussian ones. On the complete 180-series TSB-AD-M evaluation (Liu & Paparrizos, 2024), the training-free certified-box rung beats every classical and deep benchmark baseline on VUS-PR (paired CIs clear of zero), ties MMPAD, ties for first on AUC-ROC, and trails only xLSTMAD's range-aware currencies. On a commercial flight-test campaign, the certified quantile-lens rung leads the deployed operational baseline on all three currencies (ROC-AUC/VUS-ROC/VUS-PR: $0.767/0.774/0.593$ versus $0.736/0.746/0.540$).

1 INTRODUCTION

Physical systems announce their failures quietly. A chiller slowly losing refrigerant, a cooling loop fouling its heat exchanger, a turbofan in the first hours of a fan-blade defect: each drifts a few percent of operating range off its envelope long before any threshold alarm fires (Sipple, 2020; Sipple et al., 2026). Catching and evaluating these *incipient* faults from multivariate telemetry, without labeled failures, is the core problem of intelligent diagnostics: detect the outlier, attribute it to sensors (the *symptoms*), and hand the symptoms to downstream causal reasoning so a technician or engineer can act (Sipple & Youssef, 2022; Sipple et al., 2026).

A deployed and effective recipe for the detection step is *negative-sampling anomaly detection* (Sipple, 2020): label the observed telemetry as the positive class, manufacture a synthetic negative class from a reference measure over the instrument range, train a binary classifier, and read its score as an anomaly detector; Integrated Gradients (Sundararajan et al., 2017) converts each alarm into

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.

per-sensor attributions. The method lives where physical systems do, moderate dimensions, discrete operating modes, no labeled faults, and is validated in HVAC fleets and aircraft flight data (Sipple, 2020; Sipple et al., 2026). Yet it lacks sufficient theory. When does uniform sampling work, and at what dimension does it stop? Can a cleverer generator, adversarial, energy-based, boundary-seeking (Yoon et al., 2023; Grathwohl et al., 2020; Ngo et al., 2019), evade the difficulty? How many negatives, and how many *positives*, does resolving a fault of size ε cost?

This paper supplies the missing theory and the algorithms it licenses, grounded in high-dimensional probability (van Handel, 2016; Vershynin, 2018). Our contributions:

- **Limits** (§3). Under a box model of normal operation, uniform negatives avoid the decision boundary at rate exponential in D (Lemma 3.2), so every sample-dependent criterion is *blind* to boundary placement (Proposition 3.3); by a Le Cam two-point argument, *no* negative-generation scheme, uniform, learned, or adversarial, adds boundary information: negatives are ancillary, and $M \gtrsim \rho/\varepsilon$ positives are necessary at resolution ε and occupancy ρ (Theorem 3.4, Corollary 3.5).
- **The operable window** (§3.3). Uniform sampling is nevertheless usable in a dimension window: the informative band mass $h(D)$ is strictly unimodal with closed-form maximizer $D^\star = \ln(a/b)/(a-b)$, and in the incipient-fault limit the optimum obeys a universal $1/e$ *occupancy law*: operate where the mode fills $1/e \approx 37\%$ of its bounding box (Figure 1). Inside the window the certified budget grows linearly in $1/\varepsilon$; outside, exponentially in D (Theorem 3.6).
- **Constructions** (§4–§5). Seven building blocks turn the theory into detectors: shell-concentrated, importance-weighted samplers with a *dimension-free* band-hit rate (Theorem 4.2), an exact $O(D)$ envelope sampler with an end-to-end certificate covering estimated boxes and finite-sample draws (Theorem 4.5), data-seeded Langevin dynamics (van Handel, 2016) with dimension-free mixing (Theorem 4.3), a marginal-equalizing lens, an encoder sized by the $1/e$ law, and a conformal residual arm. A ladder of eight detectors assembles them.
- **Evidence** (§6). A direct dimension sweep, synthetic suites, TSB-AD-M (Liu & Paparrizos, 2024), and commercial flight-test data (Sipple et al., 2026) validate the predictions; bootstrap intervals state what each sample size resolves.

2 Background and Related Work

Following Sipple (2020), an *anomaly* is a state with near-zero probability under the normal-behavior process; detection is recast as classification against a synthetic contrast class.

Definition 2.1 (Negative-sampling anomaly detection)**.** Let normal operation generate i.i.d. observations in $\mathbb{R}^D$ with probability law P supported on a compact *normal region* $\mathcal{N}$, and let $U = \{u(1), \ldots, u(M)\}$ be the observed (positive, class-1) sample. Fix a margin $\delta > 0$ and draw the *negative* (class-0) sample $V = \{v(1), \ldots, v(N)\}$ i.i.d. from the reference uniform probability measure

$$\nu \;=\; \mathrm{Unif}\Big(\textstyle\prod_{d=1}^{D} \big[\min_m u_d(m) - \delta,\; \max_m u_d(m) + \delta \big] \Big), \tag{1}$$

the uniform law on the observed bounding box stretched by δ. Train a classifier $F : \mathbb{R}^D \to [0,1]$ on $(U,1), (V,0)$; the anomaly score of x is $1 - F(x)$; the *sampling ratio* $r_s = N/M$ is the principal dial. After normalization, $B = [0,1]^D$, $\nu = \mathrm{Unif}(B)$.

The construction is a cousin of noise-contrastive estimation (Gutmann & Hyvärinen, 2010; 2012) and of level-set estimation via classification (Steinwart et al., 2005): the negatives instantiate the *reference measure* defining the level set $\{p \geq \lambda\}$, and Lemma 4.1 shows the deployed level is exactly r_s. Alarms are attributed with Integrated Gradients (Sundararajan et al., 2017), validated for this detector class by Sipple & Youssef (2022) (unlike generic attention explanations, Jain & Wallace, 2019). Our analysis imports five tool families in the style of van Handel (2016): concentration of measure (Vershynin, 2018) (product volumes collapse exponentially, so uniform darts almost never land near a boundary, cf. Beyer et al., 1999); Le Cam's two-point method (Tsybakov, 2009), pricing the positives' boundary information; Markov semigroups and functional inequalities (tensorized Poincaré, Bakry–Émery (Bakry & Émery, 1985; Bakry et al., 2014), one-dimensional criteria (Bobkov, 1999; Bobkov & Götze, 1999), Holley–Stroock (Holley & Stroock, 1987)), with log-concave sampling theory calibrating what is achievable (Vempala & Wibisono, 2019; Chewi et al., 2022; Klartag & Lehec, 2022); Langevin dynamics in its stochastic-gradient form (Welling & Teh, 2011; Raginsky et al., 2017); and importance weighting with annealed schedules (Neal, 2001; Torrie & Valleau, 1977).

Against the wider field: one-class methods (Schölkopf et al., 2001; Vert & Vert, 2006; Ruff et al., 2018) fit a boundary to positives alone; isolation forests (Liu et al., 2008) score by splitting depth; likelihood and flow models (Dinh et al., 2017; Papamakarios et al., 2021; Behrmann et al., 2019) can rank out-of-distribution points *above* in-distribution ones (Nalisnick et al., 2019); reconstruction and transformer detectors (Kingma & Welling, 2014; Xu et al., 2022; Tuli et al., 2022) score by residual. Negative sampling differs in *what the negatives are for*: they define the estimand and set the criterion's efficiency, and Corollary 3.5 proves they can do nothing more; adaptive negative schemes (Yoon et al., 2023; Grathwohl et al., 2020; Hinton, 2002; Ngo et al., 2019; Madry et al., 2018; Carlini & Wagner, 2017) reallocate variance, never boundary information. Done right, that reallocation is Theorem 4.2.

3 The Limits of Uniform Negative Sampling

3.1 Setting, scarcity, and criterion blindness

Assumption 3.1 (Standing setting (S))**.** $B = [0,1]^D$; the normal region $\mathcal{N} \subseteq B$ has per-coordinate extent at most $\rho \in (0,1)$, i.e. $\mathcal{N} \subseteq R_\rho$ for some coordinate box R_ρ with all side lengths $\leq \rho$; P is supported on $\mathcal{N}$; negatives are i.i.d. $\mathrm{Unif}(B)$. For $t \geq 0$, $\mathcal{N}^t = \{x : d_\infty(x, \mathcal{N}) \leq t\}$ is the t-inflation in the d_∞ metric and $A_t = \mathcal{N}^t \setminus \mathcal{N}$ the *band* at depth t; an *ε-incipient fault* is a state at depth in $(0, \varepsilon]$.

Lemma 3.2 (Exponential scarcity of near negatives)**.** *Under (S), for any $t \geq 0$ with $\rho + 2t \leq 1$: (i) a single uniform negative satisfies $\mathbb{P}(d_\infty(v, \mathcal{N}) \leq t) \leq (\rho + 2t)^D$; (ii) $\mathbb{P}(\exists j : v_j \in \mathcal{N}^t) \leq \min\{1, N(\rho + 2t)^D\}$; (iii) if $N \leq \frac{1}{2}(\rho + 2t)^{-D}$, then with probability at least $\frac{1}{2}$* no *negative lies in $\mathcal{N}^t$.*

All negatives avoid an entire neighborhood of the normal region with probability exponentially close to one in D; on that event the data are silent about the boundary. Call a criterion *sample-dependent* if it evaluates classifiers only through the labeled sample, $\Phi(f; \mathcal{D}) = \varphi\big((x_i, y_i, f(x_i))_i\big)$: empirical risk with any loss, tree impurity, training cross-entropy; regularizers and architectures are inductive bias, not data.

Proposition 3.3 (Criterion blindness)**.** *Under (S), let $t > 0$ with $\rho + 2t \leq 1$, and let $E_t = \{V \cap A_t = \emptyset\}$, of probability at least $1 - N(\rho + 2t)^D$. On E_t, for every Borel C with $\mathcal{N} \subseteq C \subseteq \mathcal{N}^t$, the classifiers $\mathbf{1}_C$ and $\mathbf{1}_\mathcal{N}$ agree at every sample point, so every sample-dependent criterion assigns them identical values. In particular, with $t^*(N) = \frac{1}{2}\big((2N)^{-1/D} - \rho\big)_+$, the following holds with probability at least $\frac{1}{2}$: the data are consistent, with identical criterion values, with every declared normal region sandwiched between $\mathcal{N}$ and $\mathcal{N}^{t^*(N)}$, and a learner with adversarial (or unlucky) tie-breaking misses* every *incipient fault within depth $t^*(N)$.*

3.2 No negative-generation scheme creates boundary information

The next two results remove the loophole of drawing negatives differently.

Theorem 3.4 (Minimax lower bound; ancillarity of uniform negatives)**.** *Let $\rho \in (0,1)$, $\varepsilon \in (0, 1-\rho]$, $M, N \geq 1$. There are normal regions $\mathcal{N}_1, \mathcal{N}_2 \subseteq B$ of extent at most $\rho + \varepsilon$ with $\mathrm{Leb}(\mathcal{N}_1 \triangle \mathcal{N}_2) = \varepsilon\rho^{D-1}$, and normal laws $P_j = \mathrm{Unif}(\mathcal{N}_j)$, such that, with $\mathcal{L}_j = P_j^{\otimes M} \otimes \mathrm{Unif}(B)^{\otimes N}$ the law of the labeled sample under scenario j: (i) the negative block has the same law under both scenarios (it is ancillary), and $\mathrm{TV}(\mathcal{L}_1, \mathcal{L}_2) = \mathrm{TV}(P_1^{\otimes M}, P_2^{\otimes M}) \leq M\varepsilon/(\rho + \varepsilon)$; (ii) every estimator $\widehat{\mathcal{N}}(U, V)$ satisfies*

$$\max_{j \in \{1,2\}} \mathbb{E}_j\big[\mathrm{Leb}(\widehat{\mathcal{N}} \triangle \mathcal{N}_j)\big] \;\geq\; \frac{\varepsilon\rho^{D-1}}{2}\Big(1 - \frac{M\varepsilon}{\rho + \varepsilon}\Big). \tag{2}$$

If $M \leq (\rho + \varepsilon)/(2\varepsilon)$ the bound is $\frac{1}{4}\mathrm{Leb}(\mathcal{N}_1 \triangle \mathcal{N}_2)$, regardless of the number N of negatives.

Corollary 3.5 (No negative-generation scheme creates boundary information)**.** *Let the negatives be produced by* any *mechanism $V = g(U, \xi)$ that is a measurable (possibly randomized, via auxiliary randomness ξ independent of the scenario) function of the positives: uniform sampling, Langevin chains driven by an energy fitted to U (Grathwohl et al., 2020), near-manifold perturbations (Yoon et al., 2023), or otherwise. Then the law of (U, V) under scenario j is the image of $P_j^{\otimes M}$ under one fixed Markov kernel, total variation between scenarios cannot increase, and the bound equation 2*

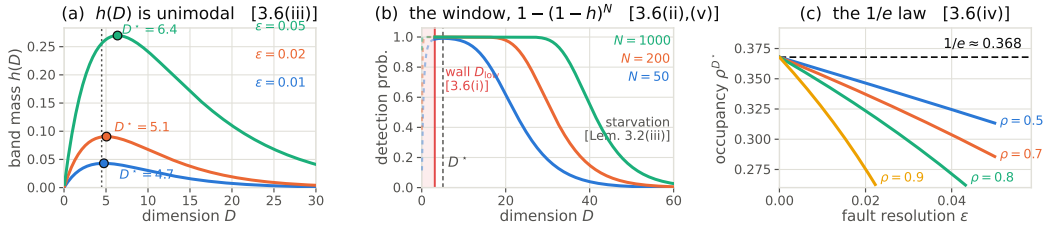


Figure 1: Theorem 3.6, drawn ($\rho = 0.8$ in a, b). *(a)* Band mass $h(D)$: unimodal, maximizer $D^\star$ (dots) sliding toward $1/\ln(1/\rho)$ as $\varepsilon \downarrow 0$. *(b)* The window: detection $1-(1-h)^N$ at $\varepsilon = 0.02$; dashed left of the wall D_{low}, starving right. *(c)* The $1/e$ law: occupancy at the optimum emanates from $e^{-1} \approx 0.368$.

holds verbatim. Consequently the positive-sample requirement $M \gtrsim \rho/\varepsilon$ for boundary resolution ε is universal over negative designs.

Corollary 3.5 does not make negatives useless; it assigns the jobs. Negatives set the *question*: they instantiate the reference measure, fixing estimand and level (Lemma 4.1). Positives supply the *answer*: they alone carry boundary information. And negatives set the *exchange rate*, from $e^{\Theta(D)}$ uniform draws per boundary hit (Lemma 3.2) to $\Theta(1)$ designed ones (Theorems 4.2 and C.3), while the destination never moves: the positives are the only witnesses to where the boundary is, the negatives the questions asked of them, and no question makes a witness know more than they saw. This is the NCE split (the noise law moves the variance, never the limit, Gutmann & Hyvärinen, 2012), so §4 optimizes negatives for pure efficiency.

3.3 The operable window and the $1/e$ law

Two walls press from opposite sides: at low D the mode fills its box and in-mode negatives corrupt the level semantics; at high D the near band starves (Lemma 3.2). Between them lies a quantifiable window (Figure 1).

Theorem 3.6 (Operable window and dimension sweet spot of uniform sampling)**.** *Assume (S) with a single* interior *box mode $\mathcal{N} = \prod_d [a_d, a_d + \rho]$, clearance $\varepsilon' = \min_d \min(a_d, 1 - a_d - \rho) > 0$, and two-sided flatness $\kappa_P \geq 1$ of P on $\mathcal{N}$ ($\kappa_P^{-1}\rho^{-D} \leq p \leq \kappa_P \rho^{-D}$). Draw $N = r_s M$ uniform negatives, fix a fault resolution $\varepsilon \in (0, \varepsilon']$ with $\rho + 2\varepsilon < 1$, and write*

$$h(D) := (\rho + 2\varepsilon)^D - \rho^D, \qquad a := \ln(1/\rho), \quad b := \ln\big(1/(\rho + 2\varepsilon)\big), \quad a > b > 0. \tag{3}$$

(i) (Lower wall.) *The estimand at the deployed level $\lambda = r_s$ is $C_{r_s} = \{p \geq r_s\}$; it equals $\mathcal{N}$ (well-posed) for all $D > D_{\text{low}} := \ln(r_s \kappa_P)/a$ and is $\emptyset$ (every normal point scored anomalous, error ρ^D) for all $D < \ln(r_s/\kappa_P)/a$.*

(ii) (Band mass and guaranteed detection.) $\text{Leb}(A_\varepsilon) = h(D)$ *exactly, and within the inflation family of Proposition 3.3 the probability that* every *sample-consistent detector flags any given fault at depth $\geq \varepsilon$ equals the band-hit probability $1 - (1 - h(D))^N$.*

(iii) (Sweet spot in closed form.) *h is strictly unimodal on $(0, \infty)$ ($\frac{dh}{dD} > 0$ before, $\frac{dh}{dD} < 0$ after), with unique maximizer*

$$D^\star = \frac{\ln(a/b)}{a - b}, \qquad \frac{1}{a} < D^\star < \frac{1}{b}, \tag{4}$$

so the guaranteed-detection probability of (ii) is maximized at $D^\star$ (integer optimum among its neighbors), and the certified budget $N^\star(D) = \lceil \ln(1/\alpha)/\ln(1/(1 - h(D))) \rceil$ at confidence $1 - \alpha$ is minimized there.

(iv) (The $1/e$ law.) *At the optimum, exactly, $h(D^\star) = \rho^{D^\star}(a - b)/b$; as $\varepsilon \downarrow 0$,*

$$D^\star \to \frac{1}{a}, \qquad \rho^{D^\star} \to e^{-1}, \qquad h(D^\star) = \frac{2\varepsilon}{e\rho a}\big(1 + O(\varepsilon)\big), \tag{5}$$

so $N^\star(D^\star) = \Theta\big(\varepsilon^{-1}\ln(1/\alpha)\big)$ *is* linear *in the resolution whereas at fixed* D *beyond the window the requirement* $N \geq \frac{1}{2}(\rho+2\varepsilon)^{-D}$ *of Lemma 3.2(iii) grows exponentially.*

(v) (Constrained optimum.) *Over the well-posed range,* $D_{\mathrm{opt}} = \max\{D^\star, D_{\mathrm{low}}\}$ *up to rounding; at* $\kappa_P = 1$ *and* $\varepsilon \downarrow 0$, $D_{\mathrm{opt}} = \max\{1, \ln r_s\}\,(1 + O(\varepsilon))/\ln(1/\rho)$, *so the sweet spot is unconstrained precisely when* $r_s < e$.

Remark 3.7 (Reading the $1/e$ law). Part (iv) says the optimum is an *occupancy*, not a dimension count: size D so the mode fills $e^{-1} \approx 37\%$ of its bounding box. The mechanism is wall area against volume: for small ε the band is a coat of paint over the mode's $2D$ faces, $h(D) \approx (2\varepsilon/\rho)D\rho^D$, and De^{-aD} peaks at $aD = 1$, one e-folding of the decay, for every ρ. With heterogeneous extents ρ_d the law is a stopping rule (grow the feature set until $\sum_d \ln(1/\rho_d) \approx 1$), and well-posedness at the deployed threshold needs $r_s\kappa_P < e$, which the default $r_s = 2$ of Sipple (2020) clears. The window's narrowness is itself the argument for §4, whose samplers remove the need to sit at $D^\star$ at all.

4 Seven Building Blocks

The theory licenses seven components; §5 composes them into detectors, and Appendix D gives a picture-first tour of all seven.

(B1) Uniform negatives. The baseline reference equation 1 of Sipple (2020). Its virtue is calibration:

Lemma 4.1 (The sampling ratio is the level)**.** *For* P *with density* p *on* B *and* $\nu = \mathrm{Unif}(B)$*, the minimizer of the surrogate risk with class weight* $\pi = M/(M+N)$ *is* $f^* = \mathbf{1}\{p \geq \lambda\}$ *with level* $\lambda = N/M = r_s$*; the excess risk of any* f *is* $\int_{\{f \neq f^*\}} |\pi p - (1-\pi)|$.

Changing the reference to any ν with density $q > 0$ while weighting negatives by $w = 1/q$ leaves every classifier's risk, hence estimand, level, and excess risk, unchanged (Lemma C.1): sampler design is importance sampling of the boundary, never a change of target.

(B2) Envelope (shell) negatives. Concentrate the reference where classifiers disagree: for a shell $S \subseteq B$ (canonically the band $A_{\varepsilon'}$) and $\gamma \in (0,1)$, the *shell mixture* is $\nu_\gamma = (1-\gamma)\mathrm{Unif}(S) + \gamma\,\mathrm{Unif}(B)$, density $q_\gamma \geq \gamma > 0$, weights $w = 1/q_\gamma$. Under the multimodal box model (M), $\mathcal{N} = \bigsqcup_{k \leq K} \mathcal{N}_k$ with sides $s_{k,i} \in [\rho_-, \rho]$, modes separated by more than $2\varepsilon'$, and $\varepsilon' \leq \delta$ (the margin of equation 1 keeps shells interior), the band-to-shell mass ratio is dimension-free (Lemma C.2), hence so is the hit rate:

Theorem 4.2 (Shell sampling hits the boundary at a dimension-free rate)**.** *Under (S) and (M) with* $S = A_{\varepsilon'}$ *and* $0 < \varepsilon \leq \varepsilon' \leq \rho_-/(2D)$*, a single draw* $v \sim \nu_\gamma$ *satisfies* $\mathbb{P}(v \in A_\varepsilon) \geq (1-\gamma)\,\frac{1}{e-1}\,\frac{\rho_-}{\rho}\,\frac{\varepsilon}{\varepsilon'}$. *One expected band sample costs* $N \leq (e-1)\frac{\varepsilon'}{\varepsilon}\frac{\rho}{\rho_-}\frac{1}{1-\gamma}$ *negatives, independent of* D*, against* $N \geq \frac{1}{2}(\rho+2\varepsilon)^{-D}$ *for uniform sampling.*

At cube modes with $\varepsilon' = 2\varepsilon$, $\gamma = \frac{1}{2}$: 7 negatives per band sample at every D, against a uniform cost past 10^{29} by $D = 100$; the same reallocation contracts the criterion's variance by a factor exponential in D (Theorem C.3). Thin shells are necessary: for $\varepsilon' \gg \rho_-/D$, volume re-concentrates in the shell's own skin.

(B3) Langevin negatives. The shell of (B2) is an oracle; sampling must be constructive, since rejection from B accepts with exponentially small probability. Under (M), the *tempered envelope* of mode k is the product law $\mu_{k,\beta} \propto e^{-\beta E_k}$, $E_k(x) = \sum_i \psi_h(\phi_{k,i}(x_i))$, with $\phi_{k,i}$ the coordinate distance to the mode and ψ_h the Huber ramp; the normalization is closed-form ($Z_k = \prod_i (s_{k,i} + 2I(\beta,h))$), so importance weights are exact. At the canonical parameters $h = 1/\beta$, $\beta = 2D/\rho_-$ the boundary layer has thickness $\rho_-/(2D)$, exactly the efficient-shell scale, and the associated Langevin diffusion mixes in dimension-free time *when seeded at the data*:

Theorem 4.3 (Box-seeded mixing in dimension-free time)**.** *Under (M) with canonical parameters, initialize the Langevin diffusion* $\mathrm{d}X_t = -\beta\nabla E_k(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ *at* $X_0 \sim \mathrm{Unif}(\mathcal{N}_k)$*. Then* $\chi^2(\mathrm{law}(X_0) \,\|\, \mu_{k,\beta}) = Z_k/\prod_i s_{k,i} - 1 \leq e^{3/2} - 1$*, and for all* $t \geq 0$*,*

$$\chi^2\big(\mathrm{law}(X_t) \,\big\|\, \mu_{k,\beta}\big) \;\leq\; e^{-2t/c_P}\,\big(e^{3/2} - 1\big), \qquad c_P(\mu_{k,\beta}) \;\leq\; e^{1/2}\big(\rho + 6/\beta\big)^2, \tag{6}$$

both bounds independent of D; the only D-dependence of the scheme is the $O(D)$ gradient cost.

The Poincaré constant comes from tensorization, a one-dimensional log-concave bound (Bobkov, 1999), and Holley–Stroock (Holley & Stroock, 1987); log-Sobolev *cannot* hold for this exponential-tailed family (Bakry et al., 2014), and is not needed. Discretized with annealed weights (Neal, 2001) under a fixed injected-variance budget, this is the C-BCLA sampler (Algorithm 2), the theory-bearing counterpart of adaptive-negative loops (Hinton, 2002; Grathwohl et al., 2020; Yoon et al., 2023). For box envelopes, however, no chain is needed at all: the product structure admits an exact $O(D)$ sampler (Proposition C.9), so Langevin only accelerates negatives for *learned* energies, priced by their oscillation from the envelope (Proposition C.10). No certificate in this paper relies on chain equilibration.

(B4) Certified boxes. Per-mode coordinate min/max boxes plus the canonical envelope give an axis-aligned estimate of each mode; a selector sweeps nested inflations $t \mapsto \mathcal{N}^t$ and accepts, at confidence $1-\alpha$, the smallest consistent with the weighted negative evidence (Algorithm 1). By Proposition 3.3 this is exactly what uniform sampling cannot supply at large D. Two statements make "certified" precise, box estimation and finite-sample draws inside the guarantee:

Definition 4.4 ((ε,α)-certificate)**.** A normal-region estimate $\widehat{\mathcal{N}}$ is (ε,α)-*certified* if, on an event of probability at least $1-\alpha$ over the training positives and the negative draws, both hold: *(coverage)* $\widehat{\mathcal{N}} \subseteq \mathcal{N}$ and every edge of $\widehat{\mathcal{N}}$ lies within ε of the corresponding edge of $\mathcal{N}$; *(resolution)* the median-of-means sweep verifies, from the weighted negatives alone, that every nested inflation of $\widehat{\mathcal{N}}$ on its grid strictly increases the empirical negative-sampling risk, so no coarser boundary is preferred by the evidence.

Theorem 4.5 (End-to-end certificate: estimated boxes, exact finite-sample draws)**.** *Assume (S) and (M) with two-sided flatness κ_P, mode separation $> 3\varepsilon'$, and $1 \le m \le D$ bands. Per mode k: estimate $\widehat{\mathcal{N}}_k$ by coordinate min/max of the M_k positives; build the envelope on $\widehat{\mathcal{N}}_k$ at canonical parameters; draw N_k negatives* exactly *from the envelope mixture, weighted by $\mathbf{1}_B/q_E$ (Proposition C.9: $O(D)$ per draw, no Markov chain, no equilibration assumption); and run the median-of-means sweep of Algorithm 1 over m inflations at grid step $\rho_-/(4Dm)$. If*

$$M_k \ge \frac{4\kappa_P \rho D^2}{\rho_-} \log \frac{4DK}{\alpha}, \qquad N_k \ge e^{1/2} \left\lceil 8 \log \tfrac{2mK}{\alpha} \right\rceil \left\lceil \tfrac{32(e-1)e^2 m(\rho/\rho_-)^2}{1-\gamma} \right\rceil,$$

then with probability at least $1-\alpha$, simultaneously for all K modes, $\widehat{\mathcal{N}} = \bigsqcup_k \widehat{\mathcal{N}}_k$ is $(\varepsilon_{\mathrm{box}},\alpha)$-certified with $\varepsilon_{\mathrm{box}} = \rho_-/(4Dm)$ (box edges are within the finer $\rho_-/(4D^2) \le \varepsilon_{\mathrm{box}}$) and the sweep returns zero inflation. The negative budget $N_k = O\big(m(\rho/\rho_-)^2 \log(mK/\alpha)\big)$ is independent of D; the total cost is $O(D(M_k+N_k))$ arithmetic per mode. Supporting results and the proof: Appendix C.4.

Definition 4.4 is what "certified" means here; box-estimation error enters as an oscillation-$\le \frac{1}{2}$ perturbation, priced by the $e^{1/2}$ in N_k (Proposition C.10). On data, CANONICAL *implements this certificate procedure*: the guarantee applies when the box, flatness, independence, partition, and budget hypotheses hold (Remark E.1).

(B5) Quantile lens. A frozen per-coordinate monotone map Φ sending each sensor's empirical marginal to $\mathrm{Unif}[0,1]$, fit on a split disjoint from certification. It equalizes every *marginal* exactly; the joint flatness improves only through the copula: under a bounded copula density $c \le C$, $\tilde{\kappa}_P \le C$, restoring the headroom $r_s \tilde{\kappa}_P < e$ when $C < e/r_s$, a working hypothesis, not a theorem. The lens also makes box modes of monotone-distorted ones, extending (M)'s reach.

(B6) Frozen encoder. A deterministic autoencoder trained on positives, then frozen, supplies a latent of width d_z; sampling, certification, and the lens run there. Theorem 3.6 makes d_z a dial: choose it so the latent occupancy satisfies $\sum_j \ln(1/\tilde{\rho}_j) \approx 1$ (the $1/e$ law), typically $d_z \in [6,12]$, the range practitioners report; representation choices (Kingma & Welling, 2014; Higgins et al., 2017; Makhzani et al., 2016; Tolstikhin et al., 2018; Kim & Mnih, 2018; Chen et al., 2018) matter only through the induced occupancy.

(B7) Residual arm. Whatever the encoder discards is invisible to every latent head, so a conformal residual detector (Vovk et al., 2005) guards the complement (a reconstruction-error radius calibrated

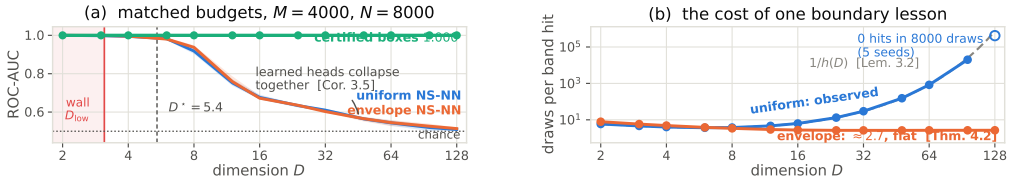


Figure 2: The headline predictions run head on: one box mode ($\rho = 0.8$), matched budgets, five seeds per D (bands are seed ranges). *(a)* Both learned heads collapse together past the window; certified boxes hold 1.000 through $D = 128$. *(b)* Draws per band hit: uniform follows the closed form $1/h(D)$ (dashed), the envelope settles at 2.6–2.7. The (a) marker is $D^\star$ at the mean depth $\bar{\varepsilon} = 0.03$; (b)'s curve uses the counted band 0.05 (optimum 6.4).

on held-out positives at budget α_f, with a visibility floor); validity holds at the level of the exchangeable calibration unit, which for telemetry is the flight or well-separated block rather than the single row. Latent detectors deploy only as unions with their residual arm.

5 The Detector Ladder and Algorithms

Table 4 (Appendix E) assembles the blocks into eight detectors; each rung changes one ingredient, so ablations are read off the ladder: `madi`, the deployed baseline of Sipple (2020); `canonical`, certified boxes for the learned head; `latent_canonical`, certification through the lens; `cbcla`, Langevin negatives; `ae_·`, heads in a frozen latent with residual arm; `cube_certify`, all seven. Procedures are Algorithms 1–5 in Appendix E; Theorem 4.5 warrants the load-bearing one.

6 Experiments

6.1 The dimension window, tested directly

We first run the paper's headline claims directly: one interior box mode ($\rho = 0.8$, $\kappa_P = 1$) in $[0,1]^D$; $M = 4{,}000$ positives, $N = 8{,}000$ negatives ($r_s = 2$); five seeds per $D \in \{2, \ldots, 128\}$; three matched arms, an NS-NN on uniform negatives (B1), the same NS-NN on exact envelope draws with weights $1/q$ (B2–B3), and certified boxes with the median-of-means selector (B4, the object of Theorem 4.5); test: 2,000 fresh normals against 2,000 incipient faults (one coordinate, depth $\mathrm{U}(0.01, 0.05)$). Figure 2: all three predictions land. *(i)* The uniform head holds through the window around $D^\star \approx 5.4$ and collapses toward chance beyond it (0.918 at $D = 8$, 0.564 at $D = 48$, 0.509 at $D = 128$); its band hits trace the unimodal band mass of Theorem 3.6(iii), and its cost per boundary lesson explodes exactly as $1/h(D)$ (Lemma 3.2): zero hits in 8,000 draws by $D \geq 96$. *(ii)* The envelope sampler's cost stays within 2.6–7.9 draws per hit across the whole grid, settling at 2.6–2.7 from $D \geq 12$ (Theorem 4.2), yet its learned head collapses *with* the uniform one (mean AUC gap ≤ 0.02 at every $D \geq 12$): ancillarity made visible, better negatives cut sampling cost but add no boundary information (Corollary 3.5). *(iii)* Certified boxes, spending the same draws on the estimand instead of a classifier head, hold ROC-AUC 1.000 at every D; the certificate *procedure* of Algorithm 1 selected zero inflation in 65/65 runs, empirical behavior at budgets below Theorem 4.5's sufficient bounds (Appendix F quantifies both and lists every dial).

6.2 Synthetic suites: mode count and mode shape

We next evaluate the eight pointwise rungs where the support is known exactly: axis-aligned hypercube modes (Assumption (M), $\kappa_P = 1$: the certified box *is* the truth) and rotated, correlated, truncated Gaussians (effective rank $\approx D/2$: any axis-aligned box is loose), crossed with $K = 1..5$ modes at $D = 16$ and three anomaly families (incipient band exits, structure violations, uniform background). Appendix G has generators and dials; Figures 10–11 plot the results.

Table 1: TSB-AD-M: the complete official evaluation list, all 180 multivariate series (TSB-AD-U unused). Above the rule our measured point-detector rungs; below, the two strongest non-pretrained leaderboard methods (official per-series results, identical series) and the benchmark's strongest classical/deep baselines, which publish per-series VUS-PR only (—). Best per column in bold; 95% CIs, paired comparisons, and the 15-series preview in Appendix H.

Method	VUS-PR	VUS-ROC	AUC-ROC
CANONICAL (certified boxes, no network)	0.359	0.765	**0.745**
AE_LATENT (ProjectionCertify)	0.348	0.751	0.725
DELAY_LATENT (delay-embedding lens)	0.346	0.771	0.740
CBCLA	0.337	0.707	0.680
MADI3 (uniform negatives; Sipple, 2020)	0.335	0.736	0.710
MADI3_ENHANCED (envelope negatives)	0.332	0.736	0.711
CUBE_CERTIFY	0.332	0.754	0.727
AE_CBCLA	0.332	0.739	0.712
LATENT_CANONICAL (quantile lens)	0.311	0.748	0.730
RESID_LATENT (innovation lens)	0.294	0.703	0.676
xLSTMAD	**0.392**	**0.778**	0.739
MMPAD	0.354	0.759	0.744
CNN	0.313	—	—
OmniAnomaly	0.312	—	—
Sub-PCA	0.310	—	—
LSTMAD	0.307	—	—

Mode count. With the dimension axis covered above, we pin $D = 16$ and sweep K: uniform-negative heads degrade with K in both geometries (MADI3 $0.92 \to 0.85$ on cubes, $1.00 \to 0.86$ on Gaussians), while rungs whose guarantees do not route through uniform occupancy stay flat: certified boxes at 1.000 for *every* K on cubes, frozen-encoder rungs at 0.97–0.99 for $K \geq 2$ on Gaussians.

Mode shape picks the rung. On cubes, CANONICAL is exact (1.000 pooled at every K) while every learned head sits at 0.82–0.92: when the box is already the right model, an encoder can only blur it. Rotated Gaussians invert the ranking: box rungs stall at 0.79 pooled, and at exactly 0.500 on the structure family (per-coordinate shuffles preserve every marginal: axis-aligned blindness made flesh), while the frozen-encoder rungs read the rotated correlations at 0.97–0.98. Background anomalies are 1.00 for every rung (all signal is near-boundary), and the incipient band on cubes separates the certified rungs (1.000/0.971 against ≤ 0.71): Theorem 4.2's band detection, delivered only by the machinery built for it. No single head wins both shapes; composition is the point, and the training-free CANONICAL reaches 0.895 grand-mean at $\sim$2 s per trial against 0.917 at $\sim$35 s for the best learned rung.

6.3 TSB-AD-M BENCHMARK

We evaluate the full ladder on the complete official TSB-AD-M evaluation list, 180 multivariate series from 17 dataset families (Liu & Paparrizos, 2024), under the official protocol: train on the anomaly-free prefix encoded in each filename, score *every* timestep, and report its measures, AUC-ROC and the range-aware VUS-ROC and VUS-PR (its headline measure), via a re-implementation of the official evaluator validated to 10^{-8}. All $1{,}800$ detector$\times$series runs of the seeded campaign completed, every dial frozen before results were read; a 15-series stratified preview, reproduced to within 0.005, carries the per-series statistics. Appendix H documents both runs and the per-family resolution (Figures 12–13).

One structural point: **every rung is a point detector**. After a causal, train-prefix-decided channel-typing pass (Appendix H), each timestep is scored as a static vector by Sections 4–5; the final score is a causal rolling maximum over one ACF period. No rung forecasts or trains on windows (the temporal rungs are lens moves feeding the same stack); the leaderboard's strongest methods are windowed or predictive by construction. Table 1 compares like with like: reference rows come from the official per-series files on the identical 180 series.

Three observations, all at $n = 180$ with paired bootstrap CIs (Appendix H). *(i) The headline measure.* On VUS-PR the training-free CANONICAL rung reaches 0.359: above every classical and deep baseline of the benchmark with paired intervals clear of zero (CNN $+0.046$, CI $[+0.008, +0.083]$; OmniAnomaly, Sub-PCA, LSTMAD $+0.047$–$+0.053$), tied with MMPAD ($+0.005$, CI $[-0.040, +0.050]$), and behind xLSTMAD (-0.033, CI $[-0.059, -0.008]$), whose lead is now resolved. On plain AUC-ROC it is statistically tied for first (0.745 against MMPAD's 0.744

Table 2: Flight-test campaign: mean per-flight scores over the 63 seeded-failure flights; columns: the eight rungs in ladder order, † deployed baseline. ROC chance 0.5; VUS-PR chance equals the anomaly rate (0.003–0.70 by mechanism); per-flight CIs in Appendix I. Best in bold.

	MADI3†	M3_ENH.	CANON.	LAT. CAN.	CBCLA	AE_LAT.	AE_CBCLA	CUBE
AUC-ROC	0.736	0.598	0.759	**0.767**	0.680	0.715	0.741	0.723
VUS-ROC	0.746	0.608	0.769	**0.774**	0.693	0.725	0.750	0.732
VUS-PR	0.540	0.443	0.547	**0.593**	0.492	0.517	0.525	0.523

and xLSTMAD's 0.739; both paired intervals cover zero), at 0.6 s of fitting per series (the certificate procedure of Algorithm 1 throughout). *(ii) Anomaly shape decides.* Split by labeled anomaly length, the long tercile belongs to the certified geometry (CANONICAL 0.855 AUC against xLSTMAD's 0.726), the short tercile to temporal modeling (0.69 best-of-ladder against 0.747): sustained faults leave $\mathrm{supp}(P)$, short in-range transients do not, and there DELAY_LATENT is xLSTMAD's closest per-series rival (mean VUS-PR rank 5.27 against 5.20). *(iii) The intervals were right, and so was the theory.* The preview's indicative VUS-ROC lead reversed at scale inside its interval (AE_LATENT − xLSTMAD -0.027, CI $[-0.053, -0.001]$): xLSTMAD keeps the range-aware currencies. The uniform- and envelope-negative heads land 0.0004 AUC apart over 180 series: Corollary 3.5 as a measurement. The certified-box machinery matches sequence models on ranking without a forecaster; the range-aware gap is the price of pointwise scoring, and the temporal lenses close most of it.

6.4 AIRCRAFT FLIGHT-TEST DATA

Our third suite is the flight-test campaign of Sipple et al. (2026): 1 Hz telemetry from the public NASA DASHlink Tail 6X regional-jet fleet, into which six failure mechanisms (fan blade out, compressor stall, progressive vibration, sensor hunting, HPT coolant loss, heat-exchanger fouling) are seeded as deterministic temporal-window perturbations of held-out flights. The deployed reference is MADI3, the multi-component negative-sampling network currently used in operations to detect and diagnose faults in piston and jet aircraft; it consumes 41 physical sensors plus spectral auxiliaries (72 inputs), while the certified rungs box the 41 physical sensors alone. All eight detectors train once on a 336-flight nominal corpus (1.69M rows) and score 75 labeled test flights (12 nominal, 63 failure) with Section 6.3's delivery and evaluator (Appendix I: every dial; Figure 15: the mechanism view).

Table 2: the quantile-lens rung LATENT_CANONICAL leads the deployed baseline on all three currencies (0.767/0.774/0.593 against 0.736/0.746/0.540; paired per-flight ΔAUC $+0.026$, 95% CI $[-0.030, +0.087]$, $n = 63$), with the training-free CANONICAL next in both ROC currencies. Mechanism resolution repeats Section 6.2: anomaly shape picks the winner. Amplitude-coded mechanisms, excursions off $\mathrm{supp}(P)$, belong to the certified geometry (fan blade out 0.989–0.997 versus 0.949; vibration 0.938–0.984 versus 0.605; compressor stall to 0.939 versus 0.702, VUS-PR 0.321 versus 0.206 at chance 0.003); the two temporally-coded mechanisms flip the ranking where spectral inputs see and raw coordinates cannot (sensor hunting, an *in-range* oscillation, 0.913 versus 0.61–0.80; HPT coolant loss 0.671 versus 0.48–0.64): the composition rule in the field. Two caveats: fouling, a drift that never leaves any sensor's band, defeats every rung alike (0.39–0.60), and the static-fit retrofit again trails its generator-trained sibling (0.598); orientation diagnostics are clean (≥ 0.98; Appendix I).

7 CONCLUSIONS AND FUTURE WORK

Negative-sampling anomaly detection now has a theory: negatives are ancillary (positives price the resolution); uniform sampling operates only in a window obeying the $1/e$ law; and Theorem 4.5 certifies the deployed object end to end, seven blocks assembling eight detectors, every rung theory-motivated, the load-bearing one certified (Remark E.1).

Limitations. The certificate is conditional on its hypotheses (k-means modes above all), and two gaps remain: deflation-side faults, untested by the inflation sweep, and attribution transfer to trained surrogates. Empirically, xLSTMAD keeps a resolved VUS-PR lead, pretrained leaderboard entries reach further still (Appendix H), the aviation lead is indicative at $n = 63$ flights, and its failures are simulator-seeded. Future work: convex-mode boundaries, attribution guarantees, and fleet-scale causal reasoning (Sipple et al., 2026).

Reproducibility statement

All experiments run end to end from the artifacts accompanying this paper. The dimension sweep is one self-contained script whose archived per-run table generates every number and figure in Section 6.1 and Appendix F. The TSB-AD-M campaign runs from one notebook against the public benchmark (resumable chunked sessions; per-series seeds fixed by position in the stratified order; one append-only result bank), and the aviation pipeline from one notebook against the public NASA DASHlink corpus with the seeded-failure generator of Sipple et al. (2026). Appendices F–I list every dial, and all appendix tables and figures are computed from the archived per-run records.

Use of large language models

The ideas and theory have evolved over a number of years, predating language models. A large language model assistant was used in preparing this work: to assist with editing, reducing, and refining the draft into the required page count (the original manuscript ran over 100 pages). We also used the language model to assist in generating the synthetic datasets, and used its deep-research mode to search for relevant related works.

REFERENCES

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, volume 1123 of *Lecture Notes in Mathematics*, pp. 177–206. Springer, 1985. URL https://doi.org/10.1007/BFb0075847.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2014. URL https://doi.org/10.1007/978-3-319-00227-9.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning (ICML)*, pp. 573–582, 2019. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999. URL https://doi.org/10.1007/3-540-49257-7_15.

Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999. URL https://doi.org/10.1214/aop/1022874820.

Sergey G. Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999. URL https://doi.org/10.1006/jfan.1998.3326.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 39–57, 2017. URL https://doi.org/10.1109/SP.2017.49.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html.

Sinho Chewi, Murat A. Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, 2022. URL https://proceedings.mlr.press/v178/chewi22a.html.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1605.08803.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=Hkxzx0NtDB.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304. PMLR, 2010. URL https://proceedings.mlr.press/v9/gutmann10a.html.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012. URL https://jmlr.org/papers/v13/gutmann12a.html.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. URL https://doi.org/10.1162/089976602760128018.

Richard Holley and Daniel Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5–6):1159–1194, 1987. URL https://doi.org/10.1007/BF01011161.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 3543–3556, 2019. URL https://aclanthology.org/N19-1357/.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, pp. 2649–2658, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

Bo'az Klartag and Joseph Lehec. Bourgain's slicing problem and KLS isoperimetry up to polylog. *Geometric and Functional Analysis*, 32(5):1134–1159, 2022. URL https://doi.org/10.1007/s00039-022-00612-9.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pp. 413–422, 2008. URL https://doi.org/10.1109/ICDM.2008.17.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards A reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. URL https://github.com/TheDatumOrg/TSB-AD.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Hy0j8ecgz.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2016. URL https://arxiv.org/abs/1511.05644.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=H1xwNhCcYm.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. URL https://doi.org/10.1023/A:1008923215028.

Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence GAN: Towards better anomaly detection. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 141–148, 2019. URL https://ieeexplore.ieee.org/document/8995288.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703, 2017. URL https://proceedings.mlr.press/v65/raginsky17a.html.

Lukas Ruff, Robert Vandermeulen, Nico Görnitz, 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 (ICML)*, 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. URL https://doi.org/10.1162/089976601750264965.

John Sipple. Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9016–9025. PMLR, 2020. URL https://proceedings.mlr.press/v119/sipple20a.html.

John Sipple and Abdou Youssef. A general-purpose method for applying explainable AI for anomaly detection. arXiv:2207.11564 [cs.LG], 2022. URL https://arxiv.org/abs/2207.11564.

John Sipple, Quang Ngoc Houng Nguyen, and Abdou Youssef. Every flight, analyzed: Scaling hybrid AI with anomaly detection, explainability, and language model causal reasoning for the exabyte era of aviation. In *IEEE/AIAA Digital Avionics Systems Conference (DASC)*, 2026.

Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005. URL https://jmlr.org/papers/v6/steinwart05a.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1711.01558.

Glenn M. Torrie and John P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977. URL https://doi.org/10.1016/0021-9991(77)90121-8.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. URL https://doi.org/10.1007/b13794.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6):1201–1214, 2022. URL https://doi.org/10.14778/3514061.3514067.

Ramon van Handel. Probability in high dimension. APC 550 Lecture Notes, Princeton University, December 2016. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/65a99bb7a3115fdede20da98b08a370f-Abstract.html.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, 2018. URL https://doi.org/10.1017/9781108231596.

Régis Vert and Jean-Philippe Vert. Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research*, 7:817–854, 2006. URL https://jmlr.org/papers/v7/vert06a.html.

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

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 681–688, 2011. URL https://dl.acm.org/doi/10.5555/3104482.3104568.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=LzQQ89U1qm_.

Sangwoong Yoon, Young-Uk Jin, Yung-Kyun Noh, and Frank C. Park. Energy-based models for anomaly detection: A manifold diffusion recovery approach. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9b6d7202750e8e32cd5270eb7fc131f7-Abstract-Conference.html.

A NOTATION AND SYMBOL GLOSSARY

Table 3 collects every recurring symbol with a pointer to its definition. Conventions: sets are calligraphic or upper-case Roman, laws and densities lower-case where possible, and hatted quantities are estimates. A few letters are traditional in two roles (B, C, g, h, K, m, π, V, Φ, Q); each such row states both, and context disambiguates. Symbols local to a single experiment appendix (the synthetic generators of Appendix G, per-notebook dial names) are defined at first use there and not repeated.

Table 3: Symbol glossary, grouped by theme; the last column is the defining occurrence.

Symbol	Meaning	Defined
Spaces, sets, and measures		
D	ambient dimension: the number of features/sensors	Def. 2.1
$B = [0,1]^D$	the normalized instrument range (the domain)	Def. 2.1
$\mathcal{N}, \mathcal{N}_k$	the normal region ($\operatorname{supp} P$); its k-th mode, an axis-aligned box under (M)	Def. 2.1; §4(B2)
$\widehat{\mathcal{N}}, \widehat{\mathcal{N}}_k$	estimated normal region; per-mode coordinate min/max box	Thm. 4.5
R_ρ	a coordinate bounding box with every side length $\leq \rho$	Ass. 3.1
$\mathcal{N}^t$	the t-inflation of $\mathcal{N}$ in the d_∞ metric	Ass. 3.1
A_t, A_ε	the band at depth t, $\mathcal{N}^t \setminus \mathcal{N}$; the incipient band	Ass. 3.1
S	the shell of the aimed mixture, canonically $A_{\varepsilon'}$	§4(B2)
d_∞	the sup (ℓ_∞) distance	Ass. 3.1
Leb	Lebesgue measure (volume)	§3
TV	total-variation distance	Thm. 3.4
$\chi^2(\cdot \,\|\, \cdot)$	chi-squared divergence	Thm. 4.3
$\triangle$	symmetric difference of sets	Thm. 3.4
E_t	the empty-band event $\{V \cap A_t = \emptyset\}$	Prop. 3.3
Q	the unit d_∞ ball (in inflations); elsewhere, the law of ξ	pf. of Lem. 3.2; pf. of Cor. 3.5
Data, classifiers, and criteria		
P, p	the law of normal operation and its Lebesgue density	Def. 2.1; Lem. 4.1
P_j, $\mathcal{L}_j$	scenario-j normal law; joint law of the labeled sample under scenario j	Thm. 3.4
U, $u(m)$, M, M_k	the positive sample, its points, its size, and the per-mode size	Def. 2.1; Thm. 4.5
V, $v(j)$, N, N_k	the negative sample, its points, its size, and the per-mode size	Def. 2.1; Thm. 4.5
ν, ν_γ, ν_E	reference measure; shell mixture; envelope mixture on estimated boxes	Def. 2.1; §4(B2); App. C.4
q, q_γ, q_E	densities of the reference measures above	§4(B1)–(B2); App. C.4
w	importance weight $\mathbf{1}_B/q$: draws outside B carry weight zero	§4(B1); App. C.4
δ	the bounding-box margin of the baseline reference	Def. 2.1
$r_s = N/M$	the sampling ratio; equals the deployed density level	Def. 2.1; Lem. 4.1
λ	the density level of the estimand $\{p \geq \lambda\}$	§2
π	class weight $M/(M+N)$ (in Lemma 4.1); per-series or per-mechanism anomaly rate, the VUS-PR chance level (in §6)	Lem. 4.1; §6.3
F, f, f^*	the trained classifier; a generic classifier; the surrogate-risk minimizer	Def. 2.1; Lem. 4.1
C, C_t, C_i	candidate normal regions; nested inflation candidates (also: the copula density bound of the lens, §4(B5))	Prop. 3.3; Thm. C.12
$\Phi(f;\mathcal{D})$, φ	a sample-dependent training criterion and its evaluation map	§3
$\widehat{R}$	the empirical (importance-weighted) negative-sampling risk	Thm. C.12
$g(U,\xi)$, ξ	a negative-generation mechanism and its auxiliary randomness (also: g is the frozen encoder in Algorithm 4)	Cor. 3.5
$t^*(N)$	the blindness depth every sample-consistent detector can miss	Prop. 3.3
Z, $Z^{(i)}$	weighted (band-)indicator statistics whose means are risk differences	Thm. C.3; Thm. C.12
θ_i, σ_i^2	mean and variance of the i-th band statistic	Thm. C.12

continued on the next page

Symbol	Meaning	Defined
Geometry of the operable window		
ρ, ρ_-	largest and smallest per-coordinate mode extent (occupancy scales)	Ass. 3.1; §4(B2)
$s_{k,i}$	side length of mode k in coordinate i, $s_{k,i} \in [\rho_-, \rho]$	§4(B2)
$\tilde{\rho}_j$	latent per-coordinate extents (the occupancy meter's input)	§4(B6)
ε	fault resolution / band depth; in the selector, the sweep grid step $\rho_-/(4Dm)$	Ass. 3.1; Thm. C.12
ε'	interior clearance of the mode; shell thickness	Thm. 3.6; §4(B2)
$\varepsilon_{\mathrm{box}}$	the certificate radius of Definition 4.4	Thm. 4.5
$\bar{\varepsilon}$	mean seeded-fault depth in the dimension sweep	§6.1
$h(D)$	the band mass $(\rho + 2\varepsilon)^D - \rho^D$ (distinct from the Huber knee h below)	Thm. 3.6
a, b	$\ln(1/\rho)$ and $\ln(1/(\rho + 2\varepsilon))$, the decay rates of mode and inflated volume	Thm. 3.6
$D^\star$, D_{low}, D_{opt}	the sweet spot $\ln(a/b)/(a-b)$; the well-posedness wall; their constrained maximum	Thm. 3.6
$N^\star(D)$	the certified uniform negative budget at confidence $1-\alpha$	Thm. 3.6
κ_P, $\tilde{\kappa}_P$	two-sided flatness of P on its mode (after the lens)	Thm. 3.6; §4(B5)
K	the number of modes (also: $K(u, E)$ is the Markov kernel in the ancillarity proof)	§4(B2); pf. of Cor. 3.5
$\widehat{w}_k$	empirical mode weight used to allocate negatives	App. C.4
γ	the uniform-mixture weight of shell and envelope references	§4(B2)
α	selector / certificate confidence	§4(B4)
m	the number of nested inflation bands the selector sweeps, $1 \le m \le D$	Thm. 4.5
K_B, N_b	median-of-means block count $\lceil 8\log(m/\alpha) \rceil$ and block size	Thm. C.12
$\hat{t}$, med_i	the selected inflation; the i-th band's median of block means	Alg. 1
Envelopes, energies, and dynamics		
E_k, E_θ	the mode-k envelope energy $\sum_i \psi_h(\phi_{k,i})$; a learned energy	§4(B3); Prop. C.10
$\phi_{k,i}$	coordinate distance to mode k's interval in coordinate i	§4(B3)
ψ_h, h	the Huber ramp and its knee $h = 1/\beta$ (distinct from the band mass $h(D)$)	§4(B3)
β	inverse temperature; canonical value $2D/\rho_-$	§4(B3)
$\mu_{k,\beta}$, μ_θ	the tempered envelope law $\propto e^{-\beta E_k}$; its learned-energy perturbation	§4(B3); Prop. C.10
Z_k, $z_{k,i}$	envelope normalizers, $Z_k = \prod_i z_{k,i}$, $z_{k,i} = s_{k,i} + 2I(\beta, h)$	§4(B3)
$I(\beta, h)$	the one-sided coordinate tail integral of the envelope	§4(B3)
J	the band tail integral $\int_0^\varepsilon e^{\beta\psi_h}$ in the transfer bound	Lem. C.11
V, c, osc	the energy perturbation $\beta(E_\theta - E_k)$, its oscillation bound, and the oscillation seminorm (this V is distinct from the negative sample V)	Prop. C.10
c_P	the Poincaré constant (relaxation time) of a law	Thm. 4.3
X_t, B_t	the Langevin diffusion and its driving Brownian motion	Thm. 4.3
$\mathcal{E}$, h_t	the Dirichlet form; the relative density of $\mathrm{law}(X_t)$	pf. of Thm. 4.3
T_r, R	annealing temperatures and the number of cooling rungs	Alg. 2
B_0	the injected-variance budget of the tempered sampler	Alg. 2
ζ	the uniform floor probability, never dropped	Alg. 2
Detectors, lenses, and evaluation		
Φ, Φ_d	the quantile lens and its per-coordinate monotone maps (distinct from the criterion $\Phi(f; \mathcal{D})$)	§4(B5)
d_z	the latent width of the frozen encoder	§4(B6)
α_f	the conformal residual arm's false-alarm budget	§4(B7)
τ_f	the calibrated conformal reconstruction-error radius	§4(B7)
$\hat{\tau}$, τ	certified and nominal score thresholds of the assembled aviation rung	App. I
AUC-ROC	area under the ROC curve over the full label vector (chance 0.5)	§6.3

continued on the next page

Symbol	Meaning	Defined
VUS-ROC, VUS-PR	the benchmark's range-aware volume-under-surface ROC and precision–recall measures (VUS-PR chance $= \pi$)	§6.3
n	the number of paired evaluation units in an interval (series or flights)	§6.3
B (resamples)	the bootstrap resample count, 20,000 (distinct from the domain B)	App. H
Indices and conventions		
d; i, j; k	coordinate index; band/draw/generic indices; mode index	throughout
$(x)_+$, $\lceil\cdot\rceil$	positive part; ceiling	Prop. 3.3; Thm. C.12
$\gtrsim$, O, Θ, Ω	inequality/order up to absolute constants	throughout
$\odot$	the elementwise (Hadamard) product	App. G
$\mathbb{E}$, Var, $\mathbb{P}$	expectation, variance, probability	throughout

B Proofs for Section 3

B.1 Scarcity and criterion blindness

Proof of Lemma 3.2. $\mathcal{N} \subseteq R_\rho$ implies $\mathcal{N}^t \subseteq R_\rho + tQ$ (Q the unit d_∞-ball), a box with side lengths at most $\rho + 2t$; hence $\mathrm{Leb}(\mathcal{N}^t) \le (\rho + 2t)^D$, which is (i) since $v \sim \mathrm{Unif}(B)$ and $\mathrm{Leb}(B) = 1$. Part (ii) is the union bound and linearity of expectation; (iii) is (ii) with $N(\rho + 2t)^D \le \frac{1}{2}$. □

Proof of Proposition 3.3. On E_t each negative lies either in $\mathcal{N}$ (where $\mathbf{1}_C = \mathbf{1}_\mathcal{N} = 1$, both committing the same labeling error) or outside $\mathcal{N}^t$ (where both vanish, since $C \subseteq \mathcal{N}^t$); each positive lies in $\mathrm{supp}(P) \subseteq \mathcal{N} \subseteq C$ (both equal 1). Hence $(x_i, y_i, \mathbf{1}_C(x_i))_i = (x_i, y_i, \mathbf{1}_\mathcal{N}(x_i))_i$ and every sample-dependent φ takes the same value. The choice $t = t^*(N)$ makes $N(\rho + 2t)^D \le \frac{1}{2}$, and Lemma 3.2(iii) applies. □

B.2 The minimax bound and ancillarity

Proof of Theorem 3.4. Take $\mathcal{N}_1 = [0, \rho]^D$ and $\mathcal{N}_2 = \mathcal{N}_1 \cup A$ with the face slab $A = [0, \rho]^{D-1} \times (\rho, \rho + \varepsilon]$; both fit in B since $\rho + \varepsilon \le 1$, and $\mathrm{Leb}(A) = \varepsilon\rho^{D-1}$. Since $\mathcal{N}_1$ and A are disjoint, $\mathrm{Leb}(\mathcal{N}_2) = \rho^D + \varepsilon\rho^{D-1} = \rho^{D-1}(\rho + \varepsilon)$, so the uniform densities are the constants $p_1 = \rho^{-D}\mathbf{1}_{\mathcal{N}_1}$ and $p_2 = \rho^{-(D-1)}(\rho + \varepsilon)^{-1}\mathbf{1}_{\mathcal{N}_2}$. We evaluate $\mathrm{TV}(P_1, P_2) = \frac{1}{2}\int_B |p_1 - p_2|$ explicitly. Partition B into $\mathcal{N}_1$, A, and $B \setminus \mathcal{N}_2$; both densities vanish on the third piece. On $\mathcal{N}_1$, factoring out ρ^{-D} resolves the absolute value: $p_1 - p_2 = \rho^{-D}\big(1 - \frac{\rho}{\rho+\varepsilon}\big) = \rho^{-D}\frac{\varepsilon}{\rho+\varepsilon} > 0$ (one unit of mass on the smaller region makes p_1 the larger density); on A, $p_1 = 0 < p_2$. Each piece therefore integrates a constant over a known volume:

$$\int_{\mathcal{N}_1} |p_1 - p_2| \;=\; \rho^D \cdot \rho^{-D}\,\frac{\varepsilon}{\rho+\varepsilon} \;=\; \frac{\varepsilon}{\rho+\varepsilon}, \qquad \int_A |p_1 - p_2| \;=\; \varepsilon\rho^{D-1} \cdot \frac{1}{\rho^{D-1}(\rho+\varepsilon)} \;=\; \frac{\varepsilon}{\rho+\varepsilon},$$

and halving the sum gives $\mathrm{TV}(P_1, P_2) = \varepsilon/(\rho + \varepsilon)$. That the two contributions are equal is conservation of mass, not coincidence: both densities integrate to one, so the excess of p_1 on $\mathcal{N}_1$ must exactly balance the excess of p_2 on the slab: each side carries the probability that a single scenario-2 positive lands where the scenarios disagree, and TV is their common value. The negative block has law $\mathrm{Unif}(B)^{\otimes N}$ under both scenarios, and total variation is unchanged by tensoring with a common factor, giving the equality in (i); $\mathrm{TV}(P^{\otimes M}, \tilde{P}^{\otimes M}) \le M\,\mathrm{TV}(P, \tilde{P})$ is the standard hybrid argument (Tsybakov, 2009). For (ii), the symmetric-difference pseudometric obeys the triangle inequality, so $\mathrm{Leb}(\widehat{\mathcal{N}} \triangle \mathcal{N}_1) + \mathrm{Leb}(\widehat{\mathcal{N}} \triangle \mathcal{N}_2) \ge \mathrm{Leb}(\mathcal{N}_1 \triangle \mathcal{N}_2)$ pointwise; Le Cam's two-point method (Lemma B.1 below, with $\Delta = \varepsilon\rho^{D-1}$) converts the tiling into equation 2. □

The corollary rests on two standard tools, stated in the form we use.

Lemma B.1 (Le Cam's two-point inequality, abstract form)**.** *Let L_1, L_2 be probability laws on a common observation space and let $\ell_1, \ell_2 \ge 0$ be measurable losses with $\ell_1(z) + \ell_2(z) \ge \Delta$ at every outcome z. Then* $\max_j \int \ell_j \,\mathrm{d}L_j \ge \frac{\Delta}{2}\big(1 - \mathrm{TV}(L_1, L_2)\big)$.

Proof. The max dominates the mean; each $\mathrm{d}L_j \geq \min(\mathrm{d}L_1, \mathrm{d}L_2)$ as measures, so the mean is at least $\frac{1}{2}\int(\ell_1+\ell_2)\min(\mathrm{d}L_1, \mathrm{d}L_2) \geq \frac{\Delta}{2}\int \min(\mathrm{d}L_1, \mathrm{d}L_2)$; and integrating $\min(a,b) = \frac{1}{2}(a+b) - \frac{1}{2}|a-b|$ gives $\int \min(\mathrm{d}L_1, \mathrm{d}L_2) = 1 - \mathrm{TV}(L_1, L_2)$. □

Lemma B.2 (Data processing for total variation)**.** *For any Markov kernel K and probability measures μ, ν on its source,* $\mathrm{TV}(\mu K, \nu K) \leq \mathrm{TV}(\mu, \nu)$.

Proof. For every event E, $\big|(\mu K)(E) - (\nu K)(E)\big| = \big|\int K(x,E)\,\mathrm{d}(\mu-\nu)\big| \leq \sup_{0\leq f\leq 1}\big|\int f\,\mathrm{d}(\mu-\nu)\big| = \mathrm{TV}(\mu,\nu)$, since $x \mapsto K(x,E)$ is one particular $[0,1]$-valued measurable function; take the supremum over E. □

Proof of Corollary 3.5. Fix the mechanism: a measurable g and auxiliary randomness $\xi \sim Q$ on a space Ξ, with both g and Q the same under either scenario. For an event E of the joint observation space set $f_E(u,\xi) = \mathbf{1}_E(u, g(u,\xi))$, jointly measurable and nonnegative, and define $K(u,E) = \int_\Xi f_E(u,\xi)\,Q(\mathrm{d}\xi)$. For fixed u this is a probability measure; measurability in u is the inner-integral clause of Tonelli's theorem; so K is a Markov kernel carrying no scenario index. Since ξ is independent of U and of the scenario, $(U,\xi) \sim P_j^{\otimes M} \otimes Q$ under scenario j, and Tonelli's iterated-evaluation clause gives, for every E,

$$\mathcal{L}_j^g(E) \;=\; \int f_E\,\mathrm{d}\big(P_j^{\otimes M}\otimes Q\big) \;=\; \int K(u,E)\,P_j^{\otimes M}(\mathrm{d}u) \;=\; \big(P_j^{\otimes M}K\big)(E)\,:$$

both scenario laws pass through the same kernel, and given the positives the manufactured negatives are ancillary. (Two readings make the step transparent. Absorbing the scenario-free randomness ξ into the input leaves the *deterministic* map $\Phi(u,\xi) = (u, g(u,\xi))$, and every output event is the input event $\Phi^{-1}(E)$ in disguise: the supremum defining TV on the output side runs over a subset of the input events, which is Lemma B.2 with no computation. And the contraction below is in fact an equality: (U,V) retains U, and projecting back is itself a kernel; only "$\leq$" is needed.) By Lemma B.2 and Theorem 3.4(i), $\mathrm{TV}(\mathcal{L}_1^g, \mathcal{L}_2^g) \leq \mathrm{TV}(P_1^{\otimes M}, P_2^{\otimes M}) \leq M\varepsilon/(\rho+\varepsilon)$. The slab tiling of the proof of Theorem 3.4 is a pointwise fact about sets, valid at every data outcome whatever V contains, so Lemma B.1 applies on the joint observation space with $L_j = \mathcal{L}_j^g$ and yields equation 2 verbatim. Finally, if $M \leq (\rho+\varepsilon)/(2\varepsilon)$ the right-hand side of equation 2 is at least $\frac{1}{4}\varepsilon\rho^{D-1}$ whatever g; contrapositively, beating that risk forces $M > (\rho+\varepsilon)/(2\varepsilon) > \rho/(2\varepsilon)$. □

B.3 Proof of Theorem 3.6

Four elementary lemmas carry the calculus; throughout, h, a, b are as in equation 3 and D ranges over $(0,\infty)$, with integer conclusions recovered by comparing integer neighbors.

Lemma B.3 (Dimension calculus of box volumes)**.** *For fixed $c \in (0,1)$, $c^D = e^{D\ln c}$ is differentiable in D with $\frac{d}{dD}c^D = c^D \ln c < 0$. In particular $\frac{d}{dD}\rho^D = -a\rho^D$ and $\frac{d}{dD}(\rho+2\varepsilon)^D = -b(\rho+2\varepsilon)^D$: each volume decays in proportion to its size, at rates $a > b > 0$.*

Proof. $c = e^{\ln c}$ and $(e^x)^y = e^{xy}$ give the rewrite; the chain rule on $e^{D\ln c}$ gives the derivative, negative since $\ln c < 0$; and $\rho < \rho + 2\varepsilon$ gives $a > b$ by monotonicity of $\ln$. □

Lemma B.4 (Unimodality of the band mass)**.** $h'(D) = a\rho^D - b(\rho+2\varepsilon)^D = e^{-aD}\big(a - be^{(a-b)D}\big)$*; h is strictly increasing on $(0, D^\star)$ and strictly decreasing on $(D^\star, \infty)$, with $D^\star$ as in equation 4 its unique critical point, characterized by the balance $a\rho^{D^\star} = b(\rho+2\varepsilon)^{D^\star}$; and $1/a < D^\star < 1/b$.*

Proof. Differentiate termwise by Lemma B.3 and factor out e^{-aD}. The bracket $\varphi(D) = a - be^{(a-b)D}$ is continuous, strictly decreasing, positive at 0, and diverges to $-\infty$, so it has exactly one zero (intermediate value theorem plus strict monotonicity), which solves $e^{(a-b)D^\star} = a/b$, i.e. equation 4. For the bracket: $D^\star = (\ln a - \ln b)/(a-b)$ is the slope of the chord of $\ln$ on $[b,a]$; the mean value theorem supplies $\xi \in (b,a)$ with $D^\star = 1/\xi$, and taking reciprocals of $b < \xi < a$ reverses the order. □

Lemma B.5 (Argmax transfer). *For $N \geq 1$ and $\alpha \in (0,1)$, $P_N(D) = 1-(1-h(D))^N$ satisfies $\frac{d}{dD}P_N = N(1-h)^{N-1}h'$, so P_N inherits the strict rise–fall of h with the same critical point; and $N^\star(D) = \lceil \ln(1/\alpha)/\ln(1/(1-h(D))) \rceil$ is the least N with $P_N \geq 1-\alpha$, non-increasing in h, hence minimized on a set containing $D^\star$.*

Proof. With $M = 1-h$ (the single-draw miss probability), $P_N = 1-M^N$; the power and chain rules give $\frac{d}{dD}P_N = -NM^{N-1}M' = N(1-h)^{N-1}h'$, and the prefactor is strictly positive on $h \in [0,1)$, so signs transfer pointwise. Solving $(1-h)^N \leq \alpha$: taking logarithms preserves the inequality, multiplying by -1 reverses it into $N\ln(1/(1-h)) \geq \ln(1/\alpha)$ with both logarithms positive, and dividing yields the ceiling; $h \mapsto \ln(1/(1-h))$ is strictly increasing, so the ratio is non-increasing in h, and h is maximal exactly at $D^\star$. □

Lemma B.6 (First-order bounds). *(i) $x - x^2/2 \leq \ln(1+x) \leq x$ for $x \geq 0$; (ii) $h \leq \ln\frac{1}{1-h} \leq h+h^2$ for $0 \leq h \leq \frac{1}{2}$; (iii) $|e^u - 1| \leq e|u|$ for $|u| \leq 1$; (iv) for $|\theta_i| \leq \frac{1}{2}$: $|1/(1+\theta) - 1| \leq 2|\theta|$ and $|(1+\theta_1)(1+\theta_2)-1| \leq |\theta_1|+|\theta_2|+|\theta_1\theta_2|$, so factors $1+O(\varepsilon)$ are closed under products, reciprocals, and quotients.*

Proof. (i) $\ln(1+x) = \int_0^x \frac{dt}{1+t}$ with $1-t \leq \frac{1}{1+t} \leq 1$; integrate. (ii) $\ln\frac{1}{1-h} = \int_0^h \frac{dt}{1-t}$ with $1 \leq \frac{1}{1-t} \leq 1+2t$ on $[0,\frac{1}{2}]$; integrate. (iii) The mean value theorem on $t \mapsto e^t$: $e^u - 1 = ue^\xi$, $|\xi| \leq |u| \leq 1$. (iv) $|1+\theta| \geq \frac{1}{2}$, and expand the product. □

Proof of Theorem 3.6. (i) Off $\mathcal{N}$, $p = 0 < r_s$; on $\mathcal{N}$, $p \in [\kappa_P^{-1}\rho^{-D}, \kappa_P\rho^{-D}]$ by flatness. If $r_s\rho^D \leq \kappa_P^{-1}$ then $r_s \leq p$ on $\mathcal{N}$ and $C_{r_s} = \mathcal{N}$; if $r_s\rho^D > \kappa_P$ then $r_s > p$ everywhere and $C_{r_s} = \emptyset$, with error $\mathrm{Leb}(\mathcal{N}) = \rho^D$ for any detector faithful to the estimand (Lemma 4.1); monotonicity of $D \mapsto r_s\rho^D$ gives the thresholds.

(ii) Interiority makes the inflation exact: $\mathcal{N}^\varepsilon = \prod_d [a_d - \varepsilon, a_d + \rho + \varepsilon] \subseteq B$, so $\mathrm{Leb}(A_\varepsilon) = (\rho+2\varepsilon)^D - \rho^D = h(D)$. Within the inflation family, a candidate C_t agrees with $\mathbf{1}_\mathcal{N}$ at every sample point iff $t < s_{\min} := \min\{d_\infty(v,\mathcal{N}) : v \in V \setminus \mathcal{N}\}$ (positives lie in every member; in-mode negatives are classified identically by all). A fault at depth $\geq \varepsilon$ is excluded by every consistent member iff $s_{\min} \leq \varepsilon$, i.e. iff some negative lands in the band, which has probability $1-(1-h(D))^N$ by independence.

(iii) Lemma B.4 gives strict unimodality with maximizer and bracket equation 4; since D enters the guarantee of (ii) and the budget only through $h(D)$, Lemma B.5 transfers the maximizer to P_N (with the integer refinement) and identifies and minimizes $N^\star$.

(iv) By the law of exponents, $h(D) = e^{-aD}(e^{(a-b)D} - 1)$; at $D^\star$ the balance of Lemma B.4 gives $e^{(a-b)D^\star} = a/b$, whence the exact identity $h(D^\star) = \rho^{D^\star}(a-b)/b$. For the limit, fix ρ and let $\varepsilon \downarrow 0$; all constants below depend only on ρ, with $\alpha \leq \frac{1}{2}$ fixed. Factoring the base and using Lemma B.6(i) at $x = 2\varepsilon/\rho$: $a - b = \ln(1+2\varepsilon/\rho) = \frac{2\varepsilon}{\rho}(1+\theta_1)$ with $-\varepsilon/\rho \leq \theta_1 \leq 0$; hence $b \geq a/2$ for $\varepsilon \leq \rho a/4$ and $b = a(1+\theta_2)$, $|\theta_2| \leq 2\varepsilon/(\rho a)$. The bracket equation 4 confines $D^\star$ to an interval of width $1/b - 1/a = (a-b)/(ab) \leq 4\varepsilon/(\rho a^2)$ whose left endpoint does not move, so by the squeeze $aD^\star = 1+\theta_3$ with $0 < \theta_3 \leq 4\varepsilon/(\rho a)$; then $\rho^{D^\star} = e^{-1}e^{-\theta_3} = e^{-1}(1+\theta_4)$, $|\theta_4| \leq e\theta_3$, by Lemma B.6(iii): the $1/e$ law, approached from below. Inserting the three estimates into the exact identity and applying Lemma B.6(iv) collapses the relative errors: $h(D^\star) = \frac{2\varepsilon}{e\rho a}(1+O(\varepsilon))$, which is equation 5. For the budget, $h^\star := h(D^\star) = \Theta(\varepsilon)$ and $h^\star \leq \frac{1}{2}$ for small ε; Lemma B.6(ii) gives $\ln(1/(1-h^\star)) = h^\star(1+\theta_5)$, $0 \leq \theta_5 \leq h^\star$, so $N^\star(D^\star) = \frac{e\rho a}{2}\varepsilon^{-1}\ln(1/\alpha)(1+O(\varepsilon)) = \Theta(\varepsilon^{-1}\ln(1/\alpha))$, the ceiling's $O(1)$ being $\varepsilon^{-1}\cdot O(\varepsilon)$. Off the window: $h(D) \leq (\rho+2\varepsilon)^D = e^{-bD}$ (Lemma 3.2(i) at $t = \varepsilon$) and Bernoulli's inequality give $P_N(D) \leq Ne^{-bD}$, so $P_N \geq \frac{1}{2}$ forces $N \geq \frac{1}{2}e^{bD}$.

(v) On $(D_{\mathrm{low}}, \infty)$ the objective of (ii) is unimodal with peak $D^\star$: if $D^\star > D_{\mathrm{low}}$ the unconstrained optimum is admissible; otherwise the admissible range is downhill and the optimum presses against the wall. Substituting $D^\star = (1+O(\varepsilon))/a$ and $D_{\mathrm{low}} = \ln r_s/a$ at $\kappa_P = 1$ gives the displayed form, and $\max\{1, \ln r_s\} = 1 \iff r_s < e$. □

C Proofs for Section 4

C.1 Calibration and invariance

Proof of Lemma 4.1. Predicting 1 at x incurs conditional risk $(1-\pi)$; predicting 0 incurs $\pi p(x)$. The pointwise minimizer selects 1 iff $\pi p(x) \geq 1-\pi$, i.e. $p(x) \geq (1-\pi)/\pi = N/M = r_s$; integrating the pointwise regret $|\pi p - (1-\pi)|$ over the disagreement set gives the excess-risk identity. □

Lemma C.1 (Exact invariance under change of reference)**.** *For every classifier f and every reference ν with density $q > 0$ on B, taking $w = 1/q$: $\mathbb{E}_{v\sim\nu}[w(v)\mathbf{1}\{f(v) = 1\}] = \int_B \mathbf{1}\{f = 1\} = \mathbb{E}_{v\sim\mathrm{Unif}(B)}[\mathbf{1}\{f(v) = 1\}]$. Hence the weighted surrogate risk, its minimizer, the level, and all excess-risk quantities are unchanged for every f simultaneously.*

Proof. $\mathbb{E}_\nu[w\,\mathbf{1}\{f = 1\}] = \int (1/q)\,\mathbf{1}\{f = 1\}\,q = \int \mathbf{1}\{f = 1\}$; the rest is Lemma 4.1 applied to the common functional. □

C.2 Shell geometry, hit rate, and variance

Lemma C.2 (Band-to-shell mass ratio)**.** *Under (S) and (M), let $0 < \varepsilon \leq \varepsilon' \leq \rho_-/(2D)$. The bands A_t are disjoint unions over modes for $t \leq \varepsilon'$, and $\mathrm{Leb}(A_\varepsilon)/\mathrm{Leb}(A_{\varepsilon'}) \geq \frac{1}{e-1}\frac{\rho_-}{\rho}\frac{\varepsilon}{\varepsilon'}$; the same bound holds for each incremental band $A_{t+\varepsilon} \setminus A_t$, and $\mathrm{Leb}(A_{\varepsilon'}) \leq K\rho^D(e-1)2D\varepsilon'/\rho_-$.*

Proof. Per mode, $L_k(t) = \prod_i(s_{k,i}+2t) - \prod_i s_{k,i} = \big(\prod_i s_{k,i}\big)\big[\prod_i(1+2t/s_{k,i}) - 1\big]$. For the lower bound, $\prod_i(1+a_i) - 1 \geq \sum_i a_i$ gives $L_k(\varepsilon) \geq \big(\prod_i s_{k,i}\big)2D\varepsilon/\rho$; for the upper, $\prod_i(1+2t/s_{k,i}) - 1 \leq e^{2Dt/\rho_-} - 1 \leq (e-1)2Dt/\rho_-$, using $2Dt/\rho_- \leq 1$ and convexity of $e^x - 1$. Divide, then pass from modes to unions by the mediant inequality. The incremental version uses that $g(t) = \prod_i(1 + 2t/s_{k,i})$ has $g'(t) \geq \sum_i 2/s_{k,i} \geq 2D/\rho$; the absolute bound uses $\prod_i s_{k,i} \leq \rho^D$. □

Proof of Theorem 4.2. $\nu_\gamma(A_\varepsilon) \geq (1-\gamma)\mathrm{Leb}(A_\varepsilon)/\mathrm{Leb}(A_{\varepsilon'})$, and Lemma C.2 bounds the ratio; the uniform column is Lemma 3.2(iii). □

Theorem C.3 (Variance contraction and effective sample size)**.** *Let f, f' be classifiers with disagreement region Δ, and $Z = w(v)(\mathbf{1}\{f(v) = 1\} - \mathbf{1}\{f'(v) = 1\})$, whose mean $\theta = \int_\Delta(\mathbf{1}\{f = 1\} - \mathbf{1}\{f' = 1\})$ is reference-independent by Lemma C.1. (i) Under $\mathrm{Unif}(B)$: $\mathrm{Var}(Z) \geq \mathrm{Leb}(\Delta)(1-\mathrm{Leb}(\Delta))$. (ii) Under ν_γ with $\Delta \subseteq S$: $\mathrm{Var}(Z) \leq \mathrm{Leb}(\Delta)\mathrm{Leb}(S)/(1-\gamma)$ and $|Z| \leq \mathrm{Leb}(S)/(1-\gamma)$ a.s., so the effective-sample-size gain is at least $(1-\gamma)(1-\mathrm{Leb}(\Delta))/\mathrm{Leb}(S)$, which is exponentially large in D, since $\mathrm{Leb}(S) \leq K\rho^D(e-1)2D\varepsilon'/\rho_-$. (iii) In general, splitting Δ, shell sampling is never worse than uniform by more than the factor $1/\gamma$.*

Proof. (i) $|Z| = \mathbf{1}_\Delta$ under $w \equiv 1$, so $\mathbb{E}Z^2 = \mathrm{Leb}(\Delta)$ and $\theta^2 \leq \mathrm{Leb}(\Delta)^2$. (ii) $\mathbb{E}[Z^2] = \int_\Delta w^2 q_\gamma = \int_\Delta w$, and on S, $w \leq \mathrm{Leb}(S)/(1-\gamma)$. (iii) On $B \setminus S$, $q_\gamma = \gamma$, so $\int_{\Delta\setminus S} w = \mathrm{Leb}(\Delta \setminus S)/\gamma$. □

C.3 Envelope properties, Poincaré, and mixing

Recall the tempered envelope of §4(B3): $E_k(x) = \sum_i \psi_h(\phi_{k,i}(x_i))$ with Huber ramp $\psi_h(t) = t^2/(2h)$ on $[0,h]$ and $t - h/2$ beyond, $\mu_{k,\beta} \propto e^{-\beta E_k}$, and the closed form $Z_k = \prod_i z_{k,i}$, $z_{k,i} = s_{k,i} + 2I(\beta,h)$, $I(\beta,h) = \sqrt{\pi h/(2\beta)}\,\mathrm{erf}(\sqrt{\beta h/2}) + \beta^{-1}e^{-\beta h/2}$ (three-zone coordinate integral: flat interior of length $s_{k,i}$, half-Gaussian shoulder, exponential tail).

Lemma C.4 (Elementary envelope properties, canonical parameters)**.** *With $h = 1/\beta$ and $\beta \geq 2/\rho_-$: (i) $0 \leq t - \psi_h(t) \leq h/2$; E_k is convex and C^1, $\mu_{k,\beta}$ is log-concave, with smoothness constant β^2. (ii) $\beta I(\beta, 1/\beta) = \sqrt{\pi/2}\,\mathrm{erf}(1/\sqrt{2}) + e^{-1/2} \approx 1.4622$, and at $\beta = 2D/\rho_-$, $\prod_i s_{k,i} \leq Z_k \leq e^{3/2}\prod_i s_{k,i}$, so $\mu_{k,\beta}(\mathcal{N}_k) \geq e^{-3/2}$. (iii) For $\beta\varepsilon \leq \frac{1}{2}$, $\mu_{k,\beta}(A_\varepsilon(\mathcal{N}_k)) \geq 2D\varepsilon e^{-2}/\rho$.*

Proof. (i) is calculus ($\psi_h'(0) = 0$ cancels the kink of ϕ at the faces; the diagonal Hessian entries are $\leq 1/h$). (ii) The displayed constant is exact, and $Z_k/\prod_i s_{k,i} = \prod_i(1 + 2I/s_{k,i}) \leq e^{\sum_i 2I/s_{k,i}} \leq$

$e^{3/2}$ using $\sum_i 2I/s_{k,i} \le 3D/(\beta\rho_-) = \frac{3}{2}$. (iii) On the i-th face slab of the band, $\psi_h(t) \le t \le \varepsilon$, so $\mu_{k,\beta}(F_i) \ge 2\varepsilon e^{-\beta\varepsilon} e^{-3/2}/s_{k,i} \ge 2\varepsilon e^{-2}/\rho$; sum over the D faces. □

Lemma C.5 (The one-dimensional envelope coordinate)**.** *Fix $s \in [\rho_-, \rho]$ and $\beta \ge 2/s$; let $g_0 \propto e^{-\beta\phi}$ (uniform of length s with two-sided exponential tails of scale $1/\beta$) and $g_h \propto e^{-\beta\psi_h(\phi)}$. Then $c_P(g_0) \le 12\mathrm{Var}(g_0) \le (s + 6/\beta)^2$ and $c_P(g_h) \le e^{\beta h/2}(s + 6/\beta)^2$.*

Proof. g_0 is a mixture of a central uniform (weight s/Z, $Z = s + 2/\beta$, second moment $s^2/12$) and two exponential tails (weight $1/(\beta Z)$ each, second moment $s^2/4 + s/\beta + 2/\beta^2$ about the center), giving $\mathrm{Var}(g_0) \le s^2/12 + s/(2\beta) + 4/\beta^2$ using $Z \ge s$ and $\beta s \ge 2$. Since g_0 is log-concave on $\mathbb{R}$, Bobkov's bound (Bobkov, 1999) gives $c_P(g_0) \le 12\mathrm{Var}(g_0) \le (s + 6/\beta)^2$, again by $\beta s \ge 2$. Finally $\beta(\phi - \psi_h(\phi)) \in [0, \beta h/2]$ by Lemma C.4(i), so Holley–Stroock (Lemma C.6) converts g_0 to g_h at cost $e^{\beta h/2}$. □

Lemma C.6 (Holley–Stroock, Poincaré form)**.** *If μ satisfies a Poincaré inequality with constant c_P and $\mu' \propto e^{-V}\mu$ with $\mathrm{osc}(V) \le c$, then μ' satisfies one with constant $e^c c_P$.*

Proof. With $Z' = \int e^{-V}\,\mathrm{d}\mu$ and the two-sided density bounds $e^{-\sup V}/Z' \le \mathrm{d}\mu'/\mathrm{d}\mu \le e^{-\inf V}/Z'$: $\mathrm{Var}_{\mu'}(f) \le \int (f - \mathbb{E}_\mu f)^2\,\mathrm{d}\mu' \le \frac{e^{-\inf V}}{Z'}\mathrm{Var}_\mu(f) \le \frac{e^{-\inf V}}{Z'} c_P \int \|\nabla f\|^2\,\mathrm{d}\mu \le e^{\sup V - \inf V} c_P \int \|\nabla f\|^2\,\mathrm{d}\mu'$. □

Theorem C.7 (Dimension-free Poincaré inequality for the envelope)**.** *Under (M), for $\beta \ge 2/\rho_-$ and $h = 1/\beta$: $c_P(\mu_{k,\beta}) \le e^{1/2}(\rho + 6/\beta)^2$, independent of the dimension; the associated Langevin diffusion has relaxation time at most this constant.*

Proof. $\mu_{k,\beta}$ is a product of coordinate laws $g_h^{(i)}$ with lengths $s_{k,i} \in [\rho_-, \rho]$. Variance tensorizes, $\mathrm{Var}(f) \le \sum_i \mathbb{E}\mathrm{Var}_i(f)$ (van Handel, 2016, §2.1); applying Lemma C.5 coordinate-wise gives $c_P(\mu_{k,\beta}) \le \max_i c_P(g_h^{(i)}) \le e^{\beta h/2} \max_i (s_{k,i} + 6/\beta)^2$ with $\beta h = 1$. The relaxation-time statement is the Poincaré–ergodicity equivalence (van Handel, 2016, §2.3–2.4). □

Proof of Theorem 4.3. On $\mathcal{N}_k$ the energy vanishes, so $\mathrm{d}\mu_{k,\beta}/\,\mathrm{dLeb} = 1/Z_k$ there while $\mathrm{d}\,\mathrm{Unif}(\mathcal{N}_k)/\,\mathrm{dLeb} = 1/\prod_i s_{k,i}$; the likelihood ratio is constant on $\mathcal{N}_k$ and zero outside, giving $\chi_0^2 = Z_k/\prod_i s_{k,i} - 1 \le e^{3/2} - 1$ by Lemma C.4(ii). The decay is the standard Poincaré consequence: with h_t the density of $\mathrm{law}(X_t)$ against $\mu_{k,\beta}$, reversibility gives $\frac{\mathrm{d}}{\mathrm{d}t}\|h_t - 1\|^2_{L^2(\mu)} = -2\mathcal{E}(h_t, h_t) \le -(2/c_P)\|h_t - 1\|^2_{L^2(\mu)}$, and $\chi^2 = \|h_t - 1\|^2_{L^2(\mu)}$; Theorem C.7 bounds c_P. □

C.4 The certificate: supporting results and the proof of Theorem 4.5

Throughout, $\nu_E := (1 - \gamma)\sum_k \widehat{w}_k\,\mu_{k,\beta} + \gamma\,\mathrm{Unif}(B)$ denotes the envelope mixture built on the estimated boxes, with density q_E and weights $w(x) = \mathbf{1}\{x \in B\}/q_E(x)$: the envelope has exponential tails on all of $\mathbb{R}^D$, and the indicator makes every weighted statistic an unbiased Lebesgue integral over B, so draws outside B carry weight zero and the $\mathrm{Unif}(B)$ estimand of Lemma 4.1 never moves (Lemma C.1 applies verbatim). $\widehat{w}_k$ is the empirical mode weight, and negatives are allocated per mode using the realized $\widehat{w}_k$, so no concentration for the mode weights is needed. Under (S)'s interior clearance, every band the selector tests lies inside B, so the indicator never truncates a tested band.

Lemma C.8 (Edge accuracy of min/max boxes)**.** *Suppose $P|_k$ has two-sided flatness κ_P on $\mathcal{N}_k$: its Lebesgue density lies between κ_P^{-1} and κ_P times $1/\prod_i s_{k,i}$. Let $\widehat{\mathcal{N}}_k$ be the coordinate min/max box of the M_k mode-k positives. Then $\widehat{\mathcal{N}}_k \subseteq \mathcal{N}_k$ always, and for any $\varepsilon_{\mathrm{box}} > 0$,*

$$\mathbb{P}\Big(\max_d\{\textit{edge gaps}\} > \varepsilon_{\mathrm{box}}\Big) \;\le\; 2D\,e^{-M_k\varepsilon_{\mathrm{box}}/(\kappa_P\rho)}.$$

Proof. An edge slab of width $\varepsilon_{\mathrm{box}}$ has $P|_k$-mass at least $\kappa_P^{-1}\varepsilon_{\mathrm{box}}/s_{k,d} \ge \varepsilon_{\mathrm{box}}/(\kappa_P\rho)$; the probability that all M_k positives miss it is at most $(1 - \varepsilon_{\mathrm{box}}/(\kappa_P\rho))^{M_k} \le e^{-M_k\varepsilon_{\mathrm{box}}/(\kappa_P\rho)}$; a union bound over the $2D$ edges completes the claim. □

Proposition C.9 (Exact $O(D)$ sampler for the envelope)**.** *$\mu_{k,\beta}$ factorizes over coordinates, and each coordinate law is an explicit three-part mixture: with probability $s_{k,i}/z_{k,i}$, uniform on the interval; with probability $I(\beta,h)/z_{k,i}$ per side, a tail at distance $t>0$ with density $\propto e^{-\beta\psi_h(t)}$, itself a two-part mixture of a truncated Gaussian on $(0,h]$ and a shifted exponential on (h,∞). All pieces admit inverse-CDF sampling, so one exact draw from $\mu_{k,\beta}$ costs $O(D)$ time, with no burn-in, no bias, and no Markov chain. Langevin dynamics is therefore* optional *for box envelopes and earns its keep only for learned energies.*

Proof. Immediate from the envelope definition in §4 (B3): the density is a product, and the closed form $z_{k,i}=s_{k,i}+2I(\beta,h)$ identifies each coordinate's mixture weights and components. □

Proposition C.10 (Robustness to perturbed energies)**.** *Let E_θ be any measurable energy with $V:=\beta(E_\theta-E_k)$ satisfying $\operatorname{osc}(V)\le c$, and $\mu_\theta\propto e^{-\beta E_\theta}$. Then: (i) proposing $x\sim\mu_{k,\beta}$ (Proposition C.9) and accepting with probability $e^{-(V(x)-\inf V)}$ yields exact draws from μ_θ at expected cost $e^c\cdot O(D)$ per draw; (ii) $c_P(\mu_\theta)\le e^c c_P(\mu_{k,\beta})$ (Lemma C.6); (iii) importance weights computed against μ_θ's unnormalized density differ from envelope weights by a factor in $[e^{-c},e^c]$, so every second-moment bound below degrades by at most e^c.*

Proof. (i) Accepted proposals have density $\propto e^{-\beta E_k}e^{-(V-\inf V)}\propto e^{-\beta E_\theta}$; the acceptance rate is $\mathbb{E}_{\mu_{k,\beta}}[e^{-(V-\inf V)}]\ge e^{-c}$. (ii) is Lemma C.6. (iii) is the two-sided density-ratio bound applied inside the second-moment integrals. □

Lemma C.11 (Second-moment transfer to the envelope reference)**.** *Under (M) with canonical parameters, let $\varepsilon\le\rho_-/(4D)$ and $\Delta\subseteq A_\varepsilon(\mathcal{N}_k)$. With negatives from ν_E and weights $w=1/q_E$,*

$$\int_\Delta \frac{\mathrm{d}x}{q_E(x)} \;\le\; \min\left\{\frac{\mathrm{Leb}(\Delta)}{\gamma},\; \frac{2(e-1)e^2\,(D\varepsilon/\rho_-)\big(\prod_i s_{k,i}\big)^2}{(1-\gamma)\,\widehat{w}_k}\right\}.$$

Proof. The first branch is $q_E\ge\gamma$ on B. For the second, on the mode-k region $q_E\ge(1-\gamma)\widehat{w}_k\,e^{-\beta E_k}/Z_k$, so $\int_\Delta 1/q_E\le\frac{Z_k}{(1-\gamma)\widehat{w}_k}\int_{A_\varepsilon}e^{\beta E_k}\,\mathrm{d}x$. By separability,

$$\int_{A_\varepsilon} e^{\beta E_k}\,\mathrm{d}x=\prod_i(s_{k,i}+2J)-\prod_i s_{k,i}\le\prod_i s_{k,i}\cdot(e-1)\,\frac{2De^{1/2}\varepsilon}{\rho_-},$$

with $J=\int_0^\varepsilon e^{\beta\psi_h(t)}\,\mathrm{d}t\le\varepsilon e^{\beta\varepsilon}\le e^{1/2}\varepsilon$, where the inequality is exactly as in Lemma C.2, using $2De^{1/2}\varepsilon/\rho_-\le e^{1/2}/2\le 1$. Multiplying by $Z_k\le e^{3/2}\prod_i s_{k,i}$ (Lemma C.4(ii)) gives the second branch. □

Theorem C.12 (Dimension-free boundary resolution by median of means)**.** *Under (M) with canonical parameters and grid step $\varepsilon=\rho_-/(4Dm)$, consider the nested candidates $C_i=\mathcal{N}^{i\varepsilon}$, $i=0,\dots,m$ (the deepest, $m\varepsilon=\rho_-/(4D)$, spans the collar), with negatives drawn from ν_E by Proposition C.9 and selection by median of means: split the negatives into $K_B=\lceil 8\log(m/\alpha)\rceil$ blocks, estimate each risk difference $\widehat{R}(C_i)-\widehat{R}(C_0)$ by the median of the block means, and return C_0 iff all m medians are positive. If, per mode,*

$$N\;\ge\;\big\lceil 8\log(m/\alpha)\big\rceil\cdot\left\lceil\frac{32\,(e-1)e^2\,m\,(\rho/\rho_-)^2}{(1-\gamma)\,\widehat{w}_k}\right\rceil\;=\;O\big(m\,(\rho/\rho_-)^2\log(m/\alpha)\big),$$

then with probability at least $1-\alpha$ the selector returns C_0. Combined with Proposition C.9, the total computational cost of ε-resolving each mode's boundary is $O\big(D\,m\,(\rho/\rho_-)^2\log(m/\alpha)\big)$ arithmetic operations, polynomial throughout and free of any $e^{\Omega(D)}$ term, against uniform sampling (Proposition 3.3).

Proof. Under (S) the positives carry no mass in any band, so the empirical risks of C_i and C_0 differ only through the weighted negatives: up to the fixed class weight, $\widehat{R}(C_i)-\widehat{R}(C_0)$ is the mean of $Z^{(i)}=w(v)\,\mathbf{1}\{v\in A_{i\varepsilon}\}$, with $\theta_i=\mathbb{E}Z^{(i)}=\mathrm{Leb}(A_{i\varepsilon})$ by Lemma C.1. Because the grid step is

$\rho_-/(4Dm)$, every tested depth obeys $i\varepsilon \le m\varepsilon = \rho_-/(4D)$, so Lemma C.11 applies at $\varepsilon \to i\varepsilon$ *within its hypothesis* for every $i \le m$, giving

$$\frac{\sigma_i^2}{\theta_i^2} \;\le\; \frac{2(e-1)e^2\,(Di\varepsilon/\rho_-)\big(\prod_i s_{k,i}\big)^2/\big((1-\gamma)\widehat{w}_k\big)}{\big(2Di\varepsilon\prod_i s_{k,i}/\rho\big)^2} \;=\; \frac{2(e-1)e^2\,m\,(\rho/\rho_-)^2}{(1-\gamma)\,\widehat{w}_k\,i},$$

using $\theta_i \ge 2Di\varepsilon\prod_i s_{k,i}/\rho$ (Lemma C.2) and $2D\varepsilon = \rho_-/(2m)$; the worst case is $i = 1$. Each block of size $N_b \ge 16\sigma_1^2/\theta_1^2$ has $\mathbb{P}(\text{block mean} \le \theta_i/2) \le 4\sigma_i^2/(N_b\theta_i^2) \le \frac{1}{4}$ by Chebyshev; the median of K_B blocks errs only if at least $K_B/2$ blocks err, which by Hoeffding has probability at most $e^{-K_B/8} \le \alpha/m$; a union bound over the m comparisons completes the argument. The stated $N = K_B N_b$ uses $16 \cdot 2(e-1)e^2 m = 32(e-1)e^2 m$; the finer grid narrows each band by the factor m, whence the linear m in the budget. Median of means is essential here: the corner weights make $\sup 1/q_E$ reach $e^{\Theta(D)}$ even though the integrals above are small, which ruins range-based (Bernstein) selection but leaves variance-based selection untouched. □

Proof of Theorem 4.5. Allocate $\alpha/(2K)$ to each mode's box event and $\alpha/(2K)$ to each selection event. Lemma C.8 with the stated M_k gives all edges within $\rho_-/(4D^2)$, which is at most the certificate radius $\varepsilon_{\text{box}} = \rho_-/(4Dm)$ because $m \le D$; with $\widehat{\mathcal{N}}_k \subseteq \mathcal{N}_k$ this is the coverage clause. On that event the envelope built on $\widehat{\mathcal{N}}_k$ differs from the true-box envelope by $V = \beta(E_{\widehat{\mathcal{N}}_k} - E_{\mathcal{N}_k})$ with $0 \le V \le \beta D \cdot \rho_-/(4D^2) = \frac{1}{2}$ pointwise (ψ_h is 1-Lipschitz and the per-coordinate distances differ by at most the edge accuracy $\rho_-/(4D^2)$), so $\operatorname{osc}(V) \le \frac{1}{2}$ and by Proposition C.10(iii) every second-moment bound in Theorem C.12 degrades by at most $e^{1/2}$, absorbed by the stated N_k. Theorem C.12 (with $m \to mK$ in the confidence split) then returns zero inflation on $\widehat{\mathcal{N}}_k$, which is the resolution clause. Mode separation $> 3\varepsilon'$ keeps the estimated shells disjoint, so the per-mode arguments do not interact. Costs: sum the budgets; each negative costs $O(D)$ by Proposition C.9. □

D THE SEVEN BUILDING BLOCKS, IN PICTURES

This appendix redraws Section 4 in plain language, one block per picture. The recurring cast: the *normal region* is a box-shaped territory inside the instrument range where healthy operation lives; an *incipient fault* is a point that has drifted a small distance ε outside its walls; and the *band* is the thin layer just outside the walls where such faults live; it is the only place a detector's boundary decision is ever tested. Formal statements are cross-referenced in brackets; nothing below is required to follow the main text, and nothing below is a metaphor with no theorem behind it.

B1: uniform negatives are darts. The baseline detector plays darts. Label the real telemetry "normal," throw N darts uniformly into the instrument range, label them "anomalous," and train any classifier to separate the two. The single dial is the sampling ratio $r_s = N/M$, and Lemma 4.1 says the dial *is* the alarm threshold: the trained classifier converges to flagging exactly the states whose normal-data density falls below r_s: throw more darts per real point and the detector demands a higher standard of "usual" before staying quiet. The catch is geometry (Figure 3). In two dimensions plenty of darts land in the band, where the boundary question is decided. But the chance that a uniform dart lands anywhere near the normal region is $(\rho + 2\varepsilon)^D$, an exponential, so by thirty dimensions essentially no dart ever visits the one place its vote matters (Lemma 3.2). The darts still set the *question* (what reference "unusual" is measured against); they simply stop showing up to help answer it.

B2: shell negatives are aimed darts. If darts never land near the walls, aim at the walls. Replace the uniform reference by a mixture: with probability $1-\gamma$, draw from a thin shell hugging the current estimate of the normal region; with probability γ, throw a plain uniform dart so that no corner of the range is ever left unwatched. Each draw carries an importance weight $1/q$, the bookkeeping that converts an aimed dart back into the currency of the original uniform question, and the invariance lemma (Lemma C.1) guarantees the exchange is exact: *aiming changes who shows up, never what is being asked*. The payoff is Theorem 4.2, drawn in Figure 4b: one useful boundary dart costs about seven aimed draws *at every dimension*, where the uniform version's price is climbing past 10^{29}.

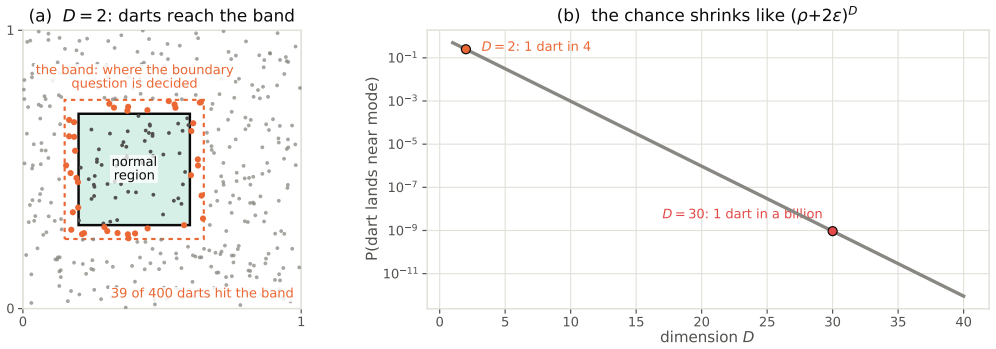


Figure 3: Block B1, uniform negatives. *(a)* At $D = 2$ the picture is friendly: of 400 uniform darts, dozens land in the band around the normal region, exactly where the boundary is decided. *(b)* The friendliness is an artifact of low dimension: the probability that a dart lands near the mode is $(\rho + 2\varepsilon)^D$, so at $D = 30$ it is roughly one dart in a billion; the training data fall silent about the boundary (Lemma 3.2, Proposition 3.3).

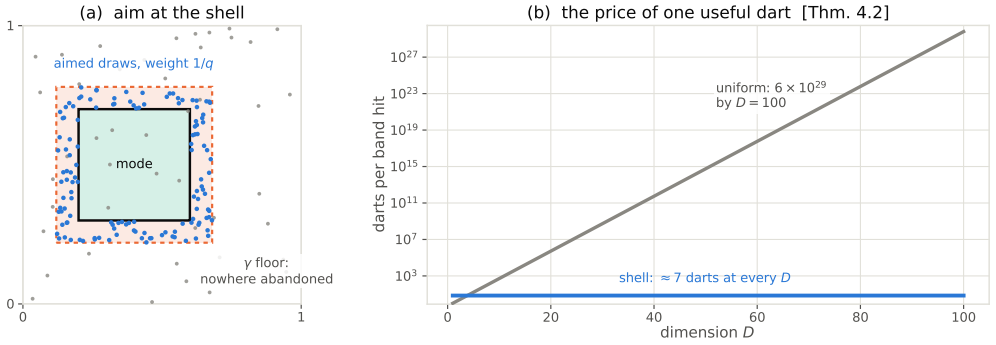


Figure 4: Block B2, envelope (shell) negatives. *(a)* The shell mixture: most draws (blue) land in a thin layer around the mode, a γ-fraction (gray) remain uniform so the reference never abandons any region; importance weights $1/q$ make the weighted sample answer the same question as uniform darts (Lemma C.1). *(b)* The exchange rate of Theorem 4.2: darts per band hit, shell versus uniform: constant against exponential.

B3: Langevin negatives are kicked marbles. A thin shell around a region is easy to want and hard to sample directly, since rejection from the full range almost never lands there (that is B1's lesson again). So manufacture the shell physically. Give each mode an energy landscape: a flat floor inside the box and a ramp that rises as you step outside. Drop marbles *at your own data points*, kick them randomly, and let them roll: kicks push them around, the ramp pushes them back, and whenever a marble pokes outside the box the ramp catches it within a thin layer along the wall, exactly the shell B2 ordered (Figure 5b). The kick size is a temperature, and the sampler cools it over rungs so early exploration is wide and the final draws hug the layer (Figure 5a). Two theorems make the marbles trustworthy: because they start at the data rather than across the huge empty range, the time to settle does not grow with dimension (Theorem 4.3); and because the landscape's normalizing constant has a closed form, each marble's importance weight is known exactly rather than estimated (Algorithm 2).

B4: certified boxes are boundaries with receipts. For each mode, per-sensor quantiles of the training data give a tight box. The open question is padding: how much slack does estimation error deserve? Sweep the padding upward in small steps and let the weighted shell darts of B2 probe each

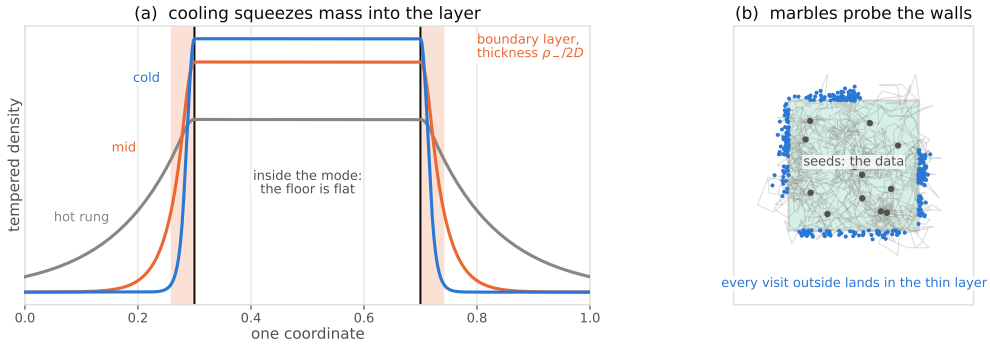


Figure 5: Block B3, data-seeded tempered Langevin negatives. *(a)* One coordinate of the tempered envelope: the floor is flat inside the mode, and as the temperature cools over rungs the exterior tail contracts into the boundary layer of thickness $\rho_-/2D$, the efficient-shell scale. *(b)* A direct simulation: marbles seeded at interior data points, kicked and cooled; every excursion outside the walls is caught by the ramp inside the thin layer, and those visits are the negatives.

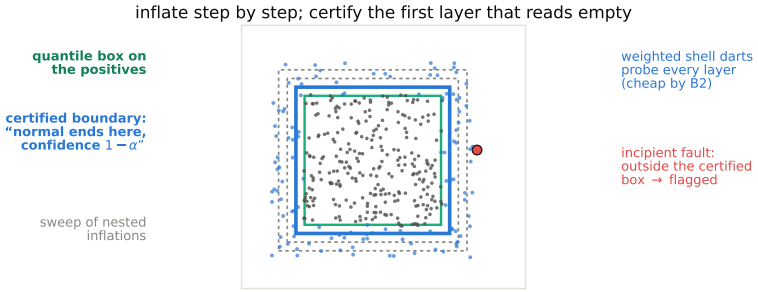


Figure 6: Block B4, certified boxes with the inflation selector (Algorithm 1). Per-sensor quantiles of the positives give the tight box; nested dashed inflations are swept outward while weighted shell darts probe each incremental layer; the first layer indistinguishable from empty at confidence $1-\alpha$ fixes the certified boundary. An incipient fault outside the certified box is flagged.

newly added layer; when a layer comes back reading empty, with no evidence that normal operation extends into it, stop and record the confidence of that emptiness verdict (Algorithm 1). The output is not just a boundary but a boundary with a receipt: "normal operation ends here, at confidence $1-\alpha$." Proposition 3.3 says such a receipt is exactly what uniform darts cannot write once the dimension grows; with shell darts the receipt's price is the fixed, dimension-free cost of Theorem 4.2. Figure 6 draws the sweep.

B5: the quantile lens flattens lumpy sensors. Real sensors are lumpy. A pressure gauge may spend ninety percent of its life between two values and make rare excursions far above; a box over the raw numbers then wastes most of its volume on readings that almost never occur, and the flatness constant κ_P of Theorem 3.6, the price of pretending the data are uniform on their box, blows up. The lens fixes each sensor separately: replace every reading by its percentile among the training data (Figure 7a). The map is monotone, so *more extreme stays more extreme* and no ordering information is lost; after it, every sensor's marginal is flat by construction, $\tilde{\kappa}_P \approx 1$, and the box machinery of B4 fits the data the way the theory assumed (Figure 7b–c). One discipline protects the guarantee: fit the lens on one half of the training data and certify boxes on the other half, so no boundary is judged by the data that drew it.

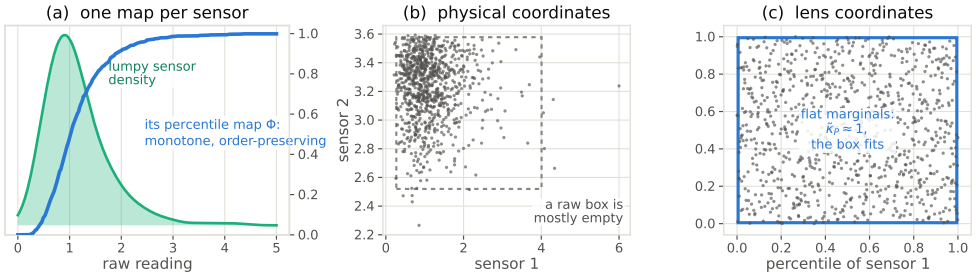


Figure 7: Block B5, the quantile lens. *(a)* One sensor's lumpy density and its percentile map Φ: monotone, hence order-preserving. *(b)* In physical coordinates a bounding box is mostly empty; the mode occupies a corner. *(c)* In lens coordinates every marginal is flat, the mode fills its box, and the certified-box machinery applies with $\tilde{\kappa}_P \approx 1$.

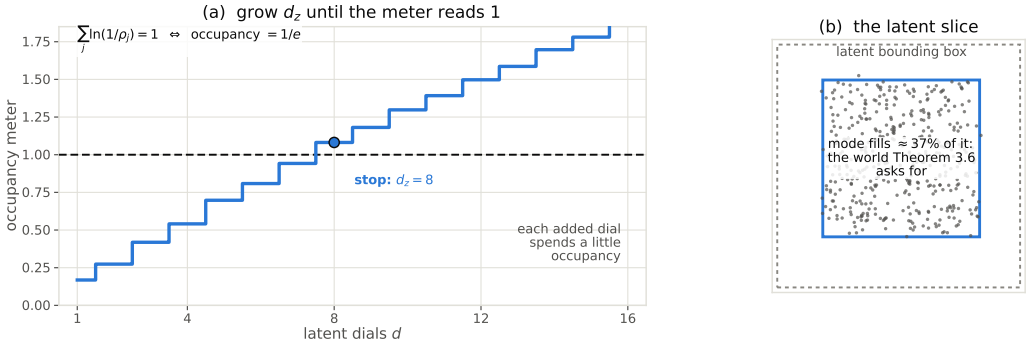


Figure 8: Block B6, the frozen encoder and the $1/e$ dial. *(a)* The stopping rule of Theorem 3.6(iv): each added latent dial spends $\ln(1/\tilde{\rho}_j)$ of occupancy; stop when the running sum crosses 1, i.e. when the mode fills $1/e$ of its latent box. *(b)* The latent slice this buys: occupancy near 37%, the geometry at which every result of Section 3.3 operates.

B6: the frozen encoder compresses to the right number of dials. Forty-one sensors are too many for any box: even at 90% occupancy per coordinate, $0.9^{41} \approx 1\%$, so the mode rattles around in a box that is ninety-nine percent empty and the band starves (Section 3). An autoencoder trained on normal data only, then *frozen* so it is a fixed lens rather than a moving target, compresses the telemetry to d_z summary dials. Theorem 3.6 turns choosing d_z from folklore into a reading on a meter: the mode should fill about $1/e \approx 37\%$ of its latent box, so add dials until the occupancy meter $\sum_j \ln(1/\tilde{\rho}_j)$ crosses 1 and stop (Figure 8a); in practice this lands at six to twelve dials. The reward is a latent world (Figure 8b) that sits exactly at the operable window's sweet spot.

B7: the residual arm guards the blind spot. Compression buys the sweet spot at a price: whatever the encoder throws away, no latent detector can see. A fault that moves the telemetry in a discarded direction produces a latent picture that looks perfectly normal. So the ladder never deploys a latent head alone. A second, deliberately simple alarm watches the *reconstruction error*, the distance between what was measured and what the encoder can express (the tube around the curve in Figure 9). Its radius is calibrated conformally (Vovk et al., 2005): rank today's error among errors on held-out normal data and alarm when it is more extreme than a $1 - \alpha_f$ fraction of them, which guarantees the false-alarm budget α_f without any distributional assumption. The deployed detector is the union: the latent head covers the map, the residual arm covers everything off the map.

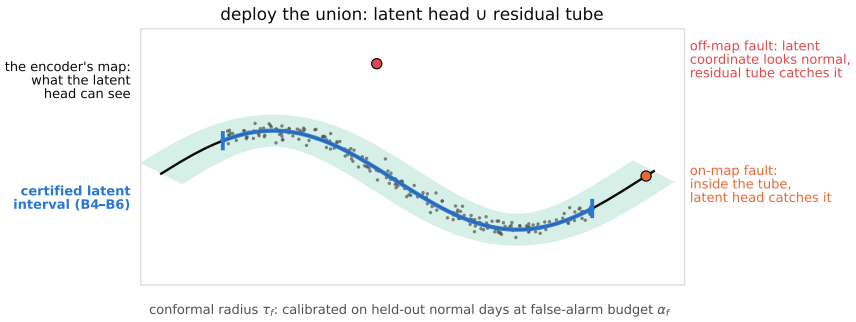


Figure 9: Block B7, the conformal residual arm. Data live near the encoder's map (curve); the tube is the conformal reconstruction-error radius τ_f, calibrated on held-out normal data at false-alarm budget α_f. The off-map fault has a perfectly normal latent coordinate and is caught only by the tube; the on-map fault is inside the tube and is caught only by the certified latent head. The ladder deploys the union of the two.

Detector	Negatives	Head	Guarantee
`madi`	uniform (B1)	NS-NN	Lem. 4.1; Thm. 3.6 window
`madi_enhanced`	envelope (B2)	NS-NN	+ Thm. 4.2 hit rate
`canonical`	envelope (B2)	certified boxes (B4)	+ finite-sample certificate
`latent_canonical`	envelope (B2)	lens (B5) + boxes (B4)	+ $\tilde{\kappa}_P \to 1$
`cbcla`	Langevin (B3)	NS-NN	+ Thm. 4.3 mixing
`ae_latent`	envelope (B2)	encoder + boxes ∪ resid. (B6, B4, B7)	+ $1/e$-law dial
`ae_cbcla`	Langevin (B3)	encoder (B6) + NS-NN ∪ resid. (B7)	+ mixing, dial
`cube_certify`	Langevin (B3)	lens + encoder + boxes ∪ resid. (B4–B7)	all of the above

Table 4: The ladder of eight detectors: each rung adds one building block (§4) and inherits the corresponding result; Remark E.1 scopes which rungs carry the formal guarantee.

E THE DETECTOR LADDER AND ADDITIONAL ALGORITHMS

Remark E.1 (The certified object and its deployed variants). Algorithm 1 exactly as stated, min/max boxes, the canonical product envelope with exact indicator-weighted draws, the median-of-means sweep at grid step $\hat{\rho}_-/(4Dm)$, is the object of Theorem 4.5. The deployed rungs harden it with theory-motivated devices that sit *outside* the formal guarantee: quantile-trimmed boxes (a contamination budget in place of min/max), the quantile lens and latent routings, score-side guards, and, in the dimension sweep, a legacy grid of m bands of width $\hat{\rho}_-/(4D)$ each. The experiment appendices state which configuration ran; "certified" rung names refer to the procedure they implement, and the finite-sample guarantee attaches when Theorem 4.5's box, flatness, independence, partition, and budget hypotheses hold.

Algorithm 1 Certified boxes with an inflation selector (blocks B2 + B4)

Require: positives U; mode count K; bands m; mixture weight γ; confidence α

1: partition U into K modes (k-means); per mode take the coordinate *min/max* box $\widehat{\mathcal{N}}_k$; set the canonical envelope $\mu_{k,\beta}$ on $\widehat{\mathcal{N}}_k$ ($\beta = 2D/\hat{\rho}_-$, $h = 1/\beta$) and the grid step $\varepsilon = \hat{\rho}_-/(4Dm)$
2: draw N negatives $v \sim \nu_E = (1-\gamma)\sum_k \widehat{w}_k \mu_{k,\beta} + \gamma\,\mathrm{Unif}(B)$ *exactly* (Proposition C.9); weight $w(v) = \mathbf{1}\{v \in B\}/q_E(v)$
3: **for** $i = 1, \dots, m$ **do**
4: $\quad \mathrm{med}_i \leftarrow$ median over $K_B = \lceil 8\log(m/\alpha) \rceil$ blocks of the block means of $w(v)\,\mathbf{1}\{v \in \widehat{\mathcal{N}}^{i\varepsilon} \setminus \widehat{\mathcal{N}}^{(i-1)\varepsilon}\}$ ▷ Theorem C.12
5: **end for**
6: **return** $\hat{t} = \varepsilon \cdot \max\big(\{0\} \cup \{\, i : \mathrm{med}_i \le 0 \,\}\big)$, with $\widehat{\mathcal{N}}^{\hat{t}}$ as the selected region; $\hat{t} = 0$, every band showing positive reference mass, is the certifying outcome of Theorem 4.5

Algorithm 2 C-BCLA: data-seeded tempered Langevin negatives (block B3, condensed)

Require: positives U (mode-partitioned boxes $\widehat{\mathcal{N}}_k$); canonical $\beta = 2D/\rho_-$, $h = 1/\beta$; rungs $T_1 > \cdots > T_R$; injected-variance budget B_0; uniform floor ζ

1: seed chains at observed positives near the mode walls, $X_0 \in U$ (box-seeding of Theorem 4.3)
2: **for** rung $r = 1, \dots, R$ **do**
3: $\quad$ run discretized Langevin steps on the tempered envelope $\mu_{k,\beta/T_r}$, spending the rung's share of B_0; carry annealed importance weights (Neal, 2001)
4: **end for**
5: mix with the uniform floor: with probability ζ emit $v \sim \mathrm{Unif}(B)$ ▷ the floor is never dropped
6: **return** weighted negatives (v, w) with the closed-form density of the envelope mixture

Algorithm 3 Quantile lens (block B5)

Require: positives U; split fraction λ_ℓ; knot count n_k

1: split U into a lens half and a certification half (disjoint)
2: per coordinate, fit a monotone quantile map Φ_d on n_k knots from the lens half; freeze $\Phi = (\Phi_1, \dots, \Phi_D)$
3: **return** Φ; downstream blocks operate on $\Phi(x)$, certification uses only the second half

Algorithm 4 Latent certified pipeline (blocks B5 + B6 + B4 ∪ B7)

Require: positives U; latent width d_z; splits for encoder / heads / conformal calibration

1: train the autoencoder on the first split; freeze the encoder g and decoder
2: choose d_z by the occupancy dial: measure latent extents $\tilde{\rho}_j$ and select d_z with $\sum_{j \le d_z} \ln(1/\tilde{\rho}_j) \approx 1$ ▷ the $1/e$ law of Theorem 3.6(iv)
3: run Algorithm 1 (optionally after Algorithm 3) in the latent on the second split
4: calibrate the conformal residual radius on the third split at false-alarm budget α_f, with a visibility floor on residual scale
5: **return** the union detector: latent alarm ∪ residual alarm

Algorithm 5 `cube_certify`: the full composition (all seven blocks)

Require: positives U; margin δ; confidences α, α_f

1: fit the quantile lens (Algorithm 3); encode with the frozen encoder at the *pinned* width $d_z = \mathrm{round}\big(1/\ln(1+2\delta)\big)$ ▷ the lens pins latent extents, so the $1/e$ dial has a closed form
2: draw Langevin negatives in the latent (Algorithm 2); train the shared NS-NN head
3: certify latent boxes with the selector (Algorithm 1) at confidence α
4: calibrate the conformal residual arm at budget α_f
5: **return** latent certified alarm ∪ residual alarm

F DIMENSION-SWEEP DETAILS

This appendix documents the sweep behind Section 6.1 and Figure 2; all 65 runs (13 dimensions $\times$ 5 seeds) completed.

Setup. The normal region is one interior box mode $\mathcal{N} = [0.1, 0.9]^D$ ($\rho = 0.8$, $\kappa_P = 1$) inside $B = [0, 1]^D$, with $D \in \{2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128\}$. Per run: $M = 4{,}000$ positives from $\mathrm{Unif}(\mathcal{N})$; the estimated box $\widehat{\mathcal{N}}$ is the coordinate min/max (the estimator of Theorem 4.5); every arm receives $N = 8{,}000$ negatives ($r_s = 2$). The uniform arm draws $\mathrm{Unif}(B)$. The envelope arm draws the mixture $\nu = (1-\gamma)\mu_\beta + \gamma\,\mathrm{Unif}(B)$ at $\gamma = \frac{1}{2}$, built on $\widehat{\mathcal{N}}$ at canonical parameters ($\beta = 2D/\hat{\rho}_-$, $h = 1/\beta$), by the exact $O(D)$ per-coordinate inverse-CDF sampler of Proposition C.9 (interior uniform with probability s_d/z_d; otherwise a truncated half-Gaussian on $(0, h]$ or a shifted exponential beyond it), with importance weights $w = 1/q_\nu$, unclipped. Both learned arms train the same NumPy MLP (64×32 ReLU, class-balanced weighted logistic loss, full-batch Adam, 600 iterations at rate 8×10^{-3}; negatives carry their importance weights inside the loss). The box arm scores collar distance to $\widehat{\mathcal{N}}$ and runs the median-of-means selector of Algorithm 1 on the same weighted draws, at the wider legacy grid predating the corrected step of Theorem C.12: $m = 8$ bands of width $\hat{\rho}_-/(4D)$ each, $K_B = \lceil 8 \ln(m/\alpha) \rceil = 41$ blocks, confidence dial $\alpha = 0.05$ (Remark E.1). Test sets: 2,000 fresh normals and 2,000 faults, each a normal draw with one uniformly chosen coordinate pushed past its wall by depth $\mathrm{U}(0.01, 0.05)$; band hits are counted against the *true* box at depth ≤ 0.05.

Reading. Sampler exactness: the uniform arm's observed draws-per-hit match $1/h(D) = (0.9^D - 0.8^D)^{-1}$ within 3% at every $D \leq 96$ (5.89 vs. 5.88 at $D = 2$; 153 vs. 158 at $D = 48$; 833 vs. 849 at $D = 64$); at $D \geq 96$ the closed form exceeds the total budget and the observed count is zero (censored marker in Figure 2b). The envelope cost sits in $[2.6, 7.9]$ draws per hit across the whole grid, the dimension-free economics of Theorem 4.2; the mild early drift reflects the canonical tail width $h = \hat{\rho}_-/(2D)$ tightening into the counted 0.05-band as D grows. The two learned heads differ by at most 0.019 mean AUC at every $D \geq 12$, and the better sampler rescues neither: the empirical face of Corollary 3.5. The selector returned $\hat{t} = 0$ (zero inflation) in 65 of 65 runs. This is the procedure's empirical selection behavior, not a per-run theorem-backed certificate: at $D = 128$ the sufficient budgets of Theorem 4.5 evaluate to $M \gtrsim 6 \times 10^5$ positives and $N \gtrsim 5 \times 10^5$ weighted draws, far above the matched $M = 4{,}000$, $N = 8{,}000$ used here, and the grid is the legacy one noted above. The theorem's constants are conservative, and the observed $65/65$ zero-inflation rate is evidence of exactly that conservatism. Separately, Figure 2a's dashed marker is $D^\star$ at the mean fault depth $\bar{\varepsilon} = 0.03$ ($D^\star = 5.42$), while panel (b)'s hit-cost curve uses the full counted band $\varepsilon = 0.05$, whose own optimum, 6.37, is visible as the minimum of the observed cost. Bands in Figure 2a are min–max over the five seeds; every number in Section 6.1 is recomputed from the archived per-run table.

G SYNTHETIC SUITE DETAILS

This appendix specifies the generators, anomaly families, architectures, and dials behind Section 6.2; Figures 10 and 11 plot the results. All 80 runs (2 geometries $\times$ $K \in \{1..5\} \times 8$ detectors) completed.

Mode placement (both geometries). K mode centers live in the unit cube $[0, 1]^{16}$: the first $\lceil \log_2 K \rceil$ coordinates carry a binary code at 0.5 ± 0.22 (any two centers differ by 0.44 somewhere), the remaining coordinates jitter uniformly in 0.5 ± 0.10. Mode extents are budgeted below half the code gap, so the separation Assumption (S) holds by construction. Mixture weights are $\mathrm{Dirichlet}(5 \cdot \mathbf{1}_K)$.

Cube geometry. Each mode is an axis-aligned product-uniform box: per-mode, per-coordinate half-widths log-uniform in $[0.05, 0.16]$ (sides 0.10–0.32, below the 0.44 gap, hence disjoint). This realizes Assumption (M) with two-sided flatness $\kappa_P = 1$: the certified boxes of Algorithm 1 coincide with the true support, so the suite has a known-optimal answer against which exactness can be read off.

Rotated-Gaussian geometry. Each mode draws $z \sim \mathcal{N}(0, I_{16})$ truncated at whitened radius $R_0 = \sqrt{16} + 2.6$ (resampling the rare exits) and maps it by $x = C_k + (z \odot s_k)Q_k^\top$: Q_k a Haar rotation (QR of a Gaussian matrix, sign-corrected), and the axis scales s_k split half "large" ($\log\mathrm{U}[0.020, 0.045]$) and half "small" ($\log\mathrm{U}[0.003, 0.010]$), randomly permuted, giving effective rank $\approx D/2$, i.e. genuine correlation structure for an encoder to find. An axis-aligned box circumscribing such a tilted ellipsoid is loose: most of its volume carries no normal mass.

The three anomaly families. Train draws 20,000 per trial; test 4,000 fresh normals (label 0) plus 1,200 anomalies (label 1) in equal thirds. *Incipient* (the band of Theorem 4.2): for cubes, a point of the assigned mode with exactly one coordinate pushed past its wall by depth $\mathrm{U}(0.01, 0.05)$; for Gaussians, a uniform direction at whitened radius $R_0 + \mathrm{U}(0.3, 1.2)$, beyond every truncation shell yet typically *inside* the circumscribing box. *Structure* (joint broken, marginals kept): for cubes, 2–3 coordinates exit together at small depths (a collar corner); for Gaussians, per-coordinate independent shuffles of real draws from the largest mode, which preserve every marginal exactly and therefore sit at chance for any axis-aligned detector. *Background*: uniform over the δ-expanded domain, rejection-sampled away from every inflated mode. Per-family AUCs in Figure 11 are computed against the same fresh normal rows.

Architectures and dials (the fairness contract). One NS-NN factory for every neural head: 64×32 ReLU with dropout 0.2, sigmoid output, Adam at 10^{-3}, ≤ 25 epochs, batch 256, patience 8 on a fixed validation mix. One autoencoder factory: mirrored 64×32 encoder/decoder with linear latent $d_z = 8$, Adam at 10^{-3}, MSE, splits $(0.5, 0.3, 0.2)$ for AE_LATENT and $(0.45, 0.40, 0.15)$ for the combined rungs, conformal residual budget $\alpha_f = 0.01$. One hardened C-BCLA curriculum for CBCLA/AE_CBCLA/CUBE_CERTIFY: 30 epochs, 20 cooling rungs $\times$ 3 ULA steps at step size 0.05, cooling factor 0.85, budget $B_0 = 0.20$, gravity 1.0, freeze skin 0.25, uniform floor $\zeta_{\max} = 0.25$ never dropped, 4,096 positives per epoch, certificate at the $q = 0.99$ energy quantile from 20,000 envelope draws at confidence 0.05. Envelope ledger: $\gamma = 0.5$, selector confidence $\alpha = 0.05$, degenerate-side floor 10^{-6}, and the hardened construction guards (trimmed mode boxes at 10^{-3} per tail; ρ_- floored at 0.02). Shared dials $r_s = 2$, $\delta = 0.05$ everywhere; each envelope pipeline is built with the trial's true K; CUBE_CERTIFY runs at its pinned width $d_z = \lceil 1/\ln(1+2\delta) \rceil = 10$. Determinism: trial seed $= 10^5 g + 10^3 K$ (g indexes the geometry) with fixed per-detector stream offsets; adaptive heads train last so no rung perturbs a predecessor's draws.

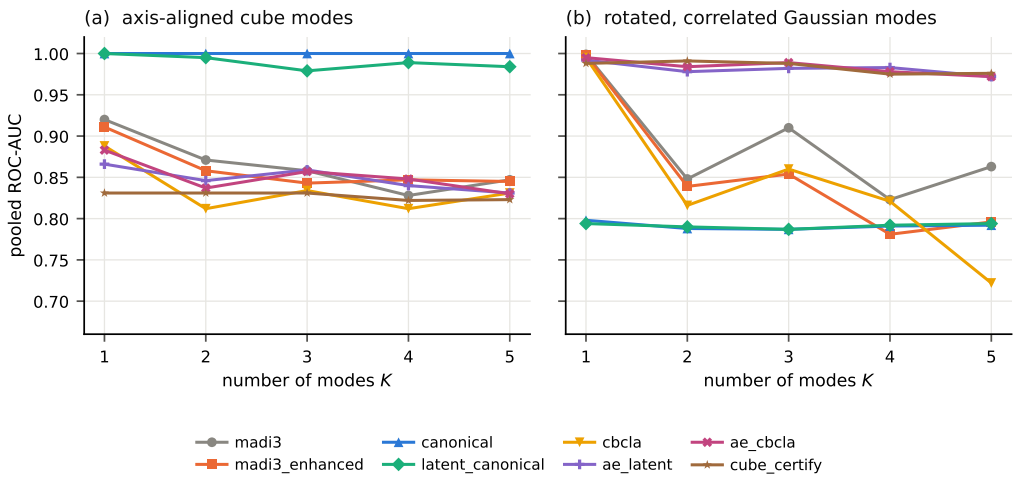


Figure 10: Pooled ROC-AUC versus the number of modes K at $D = 16$, all three anomaly families against fresh normals. *(a)* Cube modes: the certified box rung is exact at every K; the learned heads pay the mixture tax. *(b)* Rotated-Gaussian modes: the frozen-encoder rungs are flat at 0.97–0.99 while axis-aligned boxes stall near 0.79; the uniform-negative heads degrade with K in both panels.

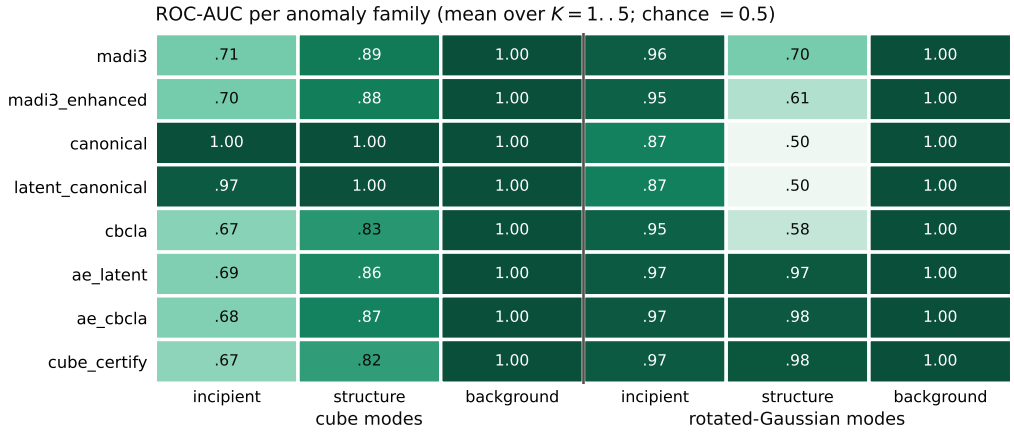


Figure 11: ROC-AUC by anomaly family (mean over K). Background is 1.00 for every rung; all the discriminative signal is near-boundary. The incipient band on cubes isolates the certified rungs (1.000/0.971 against $\leq$ 0.71); the structure family on Gaussians exposes axis-aligned blindness exactly at chance (0.50) while the encoder rungs read the broken correlation at 0.97–0.98.

H TSB-AD-M Execution Details

This appendix documents the two benchmark runs behind Section 6.3 and Table 1: the 15-series stratified preview (protocol and per-series statistics first; Figure 14 its per-family view) and the full 180-series campaign (execution, uncertainty, per-family resolution, and the per-rung analysis that closes the appendix).

Series selection and protocol (the 15-series preview). From the official TSB-AD-M evaluation list (Liu & Paparrizos, 2024) the preview takes a deterministic, family-stratified subset: families sorted alphabetically, one series per family by round-robin over the alphabetical within-family order, truncated at a compute budget of 15 series. The list has 17 families, so the rule covers one series from each of the first 15 in sort order (Genesis, MSL, Daphnet, MITDB, GHL, SMD, LTDB, SVDB, PSM, OPPORTUNITY, CreditCard, CATSv2, SMAP, GECCO, Exathlon) and the budget cuts off the last two (SWaT, TAO); the budget, the rule, and hence the series were fixed before any result was read. The selected series span $n = 1{,}827$ to 200,001 timesteps and $D = 4$ to 55 channels. The protocol is the benchmark's own: the train cutoff is parsed from the filename (`..._tr_{k}_...`); detectors train on rows $[0, k)$ only (an even row subsample capped at 30,000 rows for the pointwise rungs, the unsubsampled contiguous prefix for the two temporal rungs) and score all n rows; metrics are computed over the complete label vector. AUC-ROC is standard; VUS-ROC and VUS-PR follow the official evaluator (`generate_curve`, version `opt`, 250 thresholds) with the label-buffer window swept from 0 to one ACF period estimated by the official `find_length_rank` procedure (rank 1, first channel); our NumPy re-implementation of both the evaluator and the period estimator agrees with the reference code to 10^{-8}. Reference rows in Table 1 are recomputed from the per-series result files published in the official repository, restricted to the same 15 files; full-list reference rows in Table 1 are the means over exactly the rows of the repository's `multi_mergedTable` and `Multi_*` files (180 series); the separate univariate track (TSB-AD-U, 350 series) is not used anywhere in this paper.

The ladder under one fairness contract. All ten rungs run under shared dials: one sampling ratio $r_s = 2$, one margin $\delta = 0.05$, one NS-NN factory (64×32, dropout 0.2), one autoencoder factory (64×32, $d_z = 8$ capped at D), one hardened C-BCLA curriculum, and $K = 2$ assumed modes. Rungs therefore differ only in the ingredient the ladder varies. Every detector receives its own deterministic RNG stream derived from a per-series seed; no dial is tuned on the evaluation series. All 150 detector×series runs completed (no failures, no fallbacks invoked).

Channel typing and guards (all label-free). Before any detector sees a series, a typing pass decided on the train prefix routes each channel to the representation where its information lives:

Method	AUC-ROC	VUS-ROC	VUS-PR
CANONICAL	0.799 [0.735, 0.859]	0.809 [0.731, 0.881]	0.324 [0.189, 0.481]
AE_LATENT	0.808 [0.747, 0.869]	0.827 [0.766, 0.886]	0.318 [0.183, 0.472]
CUBE_CERTIFY	0.755 [0.685, 0.825]	0.775 [0.700, 0.849]	0.280 [0.142, 0.437]
MADI3	0.739 [0.659, 0.814]	0.762 [0.668, 0.842]	0.273 [0.137, 0.428]
LATENT_CANONICAL	0.791 [0.743, 0.842]	0.796 [0.731, 0.857]	0.272 [0.158, 0.407]
AE_CBCLA	0.740 [0.650, 0.824]	0.774 [0.682, 0.856]	0.272 [0.135, 0.428]
CBCLA	0.704 [0.609, 0.792]	0.723 [0.615, 0.818]	0.247 [0.119, 0.405]
RESID_LATENT	0.697 [0.573, 0.806]	0.728 [0.604, 0.837]	0.247 [0.150, 0.355]
DELAY_LATENT	0.723 [0.628, 0.808]	0.749 [0.652, 0.837]	0.245 [0.126, 0.389]
MADI3_ENHANCED	0.709 [0.626, 0.790]	0.724 [0.635, 0.808]	0.214 [0.106, 0.355]
xLSTMAD	0.763 [0.642, 0.859]	0.778 [0.657, 0.878]	0.350 [0.219, 0.498]
MMPAD	0.645 [0.507, 0.776]	0.656 [0.517, 0.786]	0.258 [0.127, 0.399]
CNN	—	—	0.238 [0.121, 0.384]
LSTMAD	—	—	0.225 [0.110, 0.367]
OmniAnomaly	—	—	0.212 [0.081, 0.378]
Sub-PCA	—	—	0.186 [0.100, 0.305]

Table 5: Per-method 95% bootstrap CIs on the 15-series subset ($B = 20{,}000$ series resamples). Marginal intervals overlap heavily because between-series spread dominates; the paired comparisons in the text are the sharper instrument. Classical baselines publish per-series VUS-PR only (—).

channels with exactly two training values become causal rolling event rates plus a gap-ratio column (rows since the level last changed, divided by the largest training gap, capped at 4) with three cross-channel activity summaries; near-monotone channels (net displacement ≥ 0.95 of total variation) become first differences; piecewise-constant setpoint channels ($\geq 98\%$ zero increments) become deviations from a slow running baseline (EMA, halflife $\min(k/8, 2000)$). Train-constant channels that depart on at most 5% of rows are kept with a seeded micro-dither and a certified sentinel arm folded into every score; frequent departers are dropped. Two run-time guards act on the score vector using only constructed data: an orientation check (uniform draws in the inflated train box, or injected spikes for the temporal rungs, must out-rank held-out train rows; a separation AUC below 0.35 flips the score) and a wall-clock cap with a declared fallback. Every action is recorded per run; on this subset the lenses fired on 8 of 15 series and no flip or fallback was needed.

Score delivery and ablation. The final score of every rung is passed through a causal rolling maximum over one ACF period, the same window the VUS family sweeps. This is the only post-processing step, it is monotone in the evidence, and it converts point detections into sustained window coverage. Ablation (identical pipeline, smoother off): best VUS-PR falls from 0.324 to 0.292 and best VUS-ROC from 0.827 to 0.812, while best AUC-ROC falls from 0.808 to 0.786; the smoother improves all three, i.e. it repairs delivery without buying ranking.

Statistical uncertainty. All intervals are 95% percentile bootstrap confidence intervals over series ($B = 20{,}000$ resamples of the 15 series with replacement); Table 5 lists them for every method and currency. Because every method runs on identical series, ordering claims use *paired* resampling: the series are resampled jointly and the difference of subset means is recomputed on each resample, which removes the between-series variance that the marginal intervals share. The headline comparisons: AE_LATENT $-$ xLSTMAD is $+0.049$, CI $[-0.037, +0.151]$ on VUS-ROC (fraction of bootstrap mass above zero 0.85) and $+0.045$, CI $[-0.044, +0.148]$ on AUC-ROC (0.82); CANONICAL $-$ xLSTMAD on VUS-PR is -0.027, CI $[-0.152, +0.086]$; CANONICAL $-$ MMPAD on VUS-PR is $+0.066$, CI $[-0.081, +0.197]$ (0.83); CANONICAL $-$ CNN on VUS-PR is $+0.086$, CI $[+0.028, +0.148]$ (> 0.999). Only the last interval excludes zero. At $n = 15$, then, the subset resolves the certified rungs' advantage over the classical baselines on the headline measure and leaves the orderings among the modern methods, in both directions, as point-estimate trends; this is a statement about interval width at the observed between-series spread. The full campaign below is the referee, and it confirms both readings: the classical advantage persists and resolves further, while the indicative xLSTMAD orderings resolve in the reference's favor.

The full-benchmark campaign. The 180-series evaluation ran as a resumable campaign: chunked hosted sessions append to one result bank keyed by (series, rung), each series' seed is fixed by its

Method	AUC-ROC	VUS-ROC	VUS-PR
CANONICAL	0.745 [0.718, 0.771]	0.765 [0.738, 0.792]	0.359 [0.313, 0.407]
AE_LATENT	0.725 [0.696, 0.754]	0.751 [0.722, 0.779]	0.348 [0.301, 0.398]
DELAY_LATENT	0.740 [0.712, 0.768]	0.771 [0.744, 0.797]	0.346 [0.299, 0.396]
CBCLA	0.680 [0.647, 0.712]	0.707 [0.673, 0.741]	0.337 [0.287, 0.387]
MADI3	0.710 [0.681, 0.739]	0.736 [0.705, 0.766]	0.335 [0.286, 0.385]
MADI3_ENHANCED	0.711 [0.682, 0.739]	0.736 [0.707, 0.765]	0.332 [0.284, 0.383]
CUBE_CERTIFY	0.727 [0.700, 0.754]	0.754 [0.726, 0.781]	0.332 [0.286, 0.380]
AE_CBCLA	0.712 [0.682, 0.740]	0.739 [0.709, 0.768]	0.332 [0.282, 0.383]
LATENT_CANONICAL	0.730 [0.705, 0.755]	0.748 [0.722, 0.774]	0.311 [0.273, 0.349]
RESID_LATENT	0.676 [0.643, 0.708]	0.703 [0.670, 0.735]	0.294 [0.253, 0.336]
xLSTMAD	0.739 [0.704, 0.773]	0.778 [0.744, 0.811]	0.392 [0.344, 0.442]
MMPAD	0.744 [0.712, 0.776]	0.759 [0.727, 0.790]	0.354 [0.303, 0.406]
CNN	—	—	0.313 [0.265, 0.363]
OmniAnomaly	—	—	0.312 [0.263, 0.364]
Sub-PCA	—	—	0.310 [0.258, 0.364]
LSTMAD	—	—	0.307 [0.257, 0.359]

Table 6: Full campaign: per-method means with 95% bootstrap CIs over all 180 series ($B = 20{,}000$ series resamples). References from the official per-series files on the identical series; classical baselines publish per-series VUS-PR only (—).

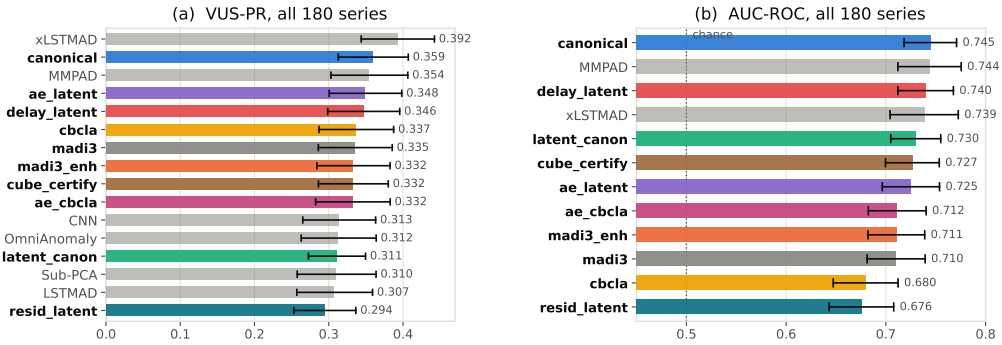


Figure 12: The full campaign at a glance: per-method means with 95% bootstrap CIs over all 180 series, our rungs in their ladder colors, references in gray. *(a)* VUS-PR: CANONICAL sits above every classical and deep baseline and every rung, tied with MMPAD, behind xLSTMAD. *(b)* AUC-ROC: CANONICAL leads on the point estimate, with MMPAD, DELAY_LATENT, and xLSTMAD within 0.006, a statistical tie.

position in the family-stratified order (so results are independent of how sessions are chunked), the tuned dial books and every guard of the preview run apply unchanged, and a banked series is never re-run. All 1,800 detector×series runs completed, with zero orientation flips and zero fallbacks (median orientation-separation AUC 0.993, 5th percentile 0.879); the three SMAP series whose training prefixes leave a single live channel route their train-constant command channels through the sentinel-and-dither lens and score normally. Total fit-and-score cost for the whole campaign: 7.6 hours on one hosted GPU runtime, from 0.6 s per series (CANONICAL) to 35 s (AE_CBCLA). Restricted to the 15 preview files, the campaign reproduces the preview's point estimates to within 0.005 (largest drift: AE_LATENT VUS-PR). Table 6 gives the marginal intervals behind Table 1; Figure 12 draws them.

Paired comparisons at $n = 180$. Joint series resampling ($B = 20{,}000$) around the difference of means. On VUS-PR, CANONICAL − CNN is $+0.046$ $[+0.008, +0.083]$, − OmniAnomaly $+0.047$ $[+0.013, +0.080]$, − Sub-PCA $+0.050$ $[+0.018, +0.081]$, − LSTMAD $+0.053$ $[+0.019, +0.086]$: every classical and deep baseline of the benchmark is now resolved in the certified rung's favor. Against the modern methods, − MMPAD is $+0.005$ $[-0.040, +0.050]$ (a statistical tie) and − xLSTMAD -0.033 $[-0.059, -0.008]$: the reference's headline lead is real. On AUC-ROC, CANONICAL

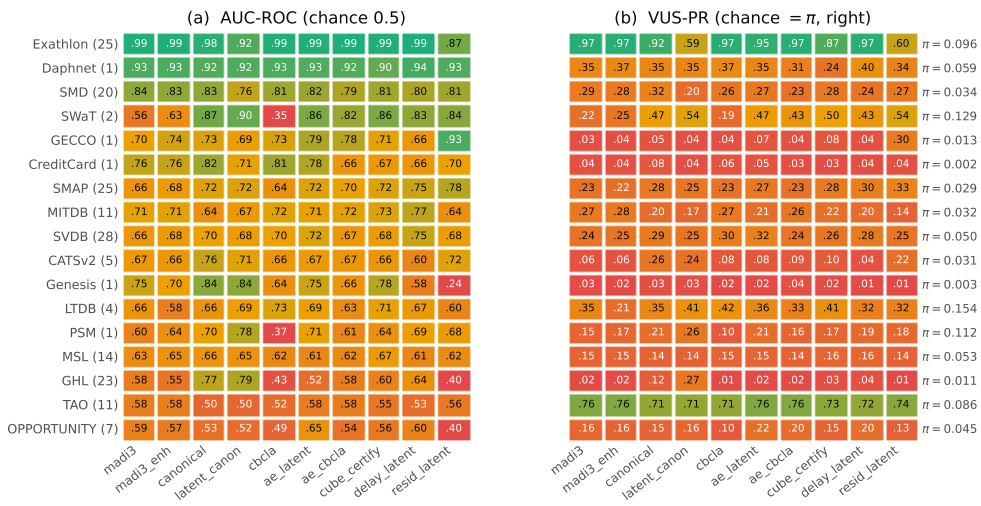


Figure 13: Full campaign, family-resolved: mean AUC-ROC *(a)* and VUS-PR *(b)* per dataset family (series counts left, per-family anomaly rates π right; rows sorted by family mean AUC). VUS-PR cells are comparable down a column only. The vertical banding is the composition rule of Section 6.2 in the wild; the per-rung reading follows below.

$-$ xLSTMAD is $+0.006$ $[-0.028, +0.040]$ and $-$ MMPAD $+0.001$ $[-0.035, +0.037]$, first place by point estimate, unresolved as an ordering. On VUS-ROC, CANONICAL $-$ xLSTMAD is -0.013 $[-0.047, +0.022]$ and DELAY_LATENT $-$ xLSTMAD -0.007 $[-0.037, +0.023]$ (unresolved), while AE_LATENT $-$ xLSTMAD, the preview's indicative $+0.049$, lands at -0.027 $[-0.053, -0.001]$: resolved for the reference, inside the preview's interval. For completeness, newer leaderboard entries built on pretraining or external corpora reach 0.361–0.495 VUS-PR on the same series (TSPulse, CHARM, StreamVAE, AxonAD); the from-scratch comparisons above are the ones this paper's theory speaks to.

What each rung is for. Figure 13 makes the composition rule quantitative; column by column: CANONICAL (certified boxes) owns sustained amplitude excursions, faults that leave the marginal support and stay out. It takes the long-anomaly tercile at 0.855 mean AUC (xLSTMAD: 0.726) and the level-coded industrial families (SWaT 0.87, CreditCard 0.82, Genesis 0.84, CATSv2 0.76) at 0.6 s per series; its weakness is any anomaly that never exits a marginal box: in-range rhythm changes (MITDB 0.64 against DELAY_LATENT's 0.77) and the short-anomaly tercile (0.68 against xLSTMAD's 0.75). LATENT_CANONICAL (quantile lens) is the drift-and-setpoint specialist: on GHL, whose faults hide beneath wandering setpoints at $\pi = 0.011$, it posts 0.79 AUC and 0.27 VUS-PR, an order of magnitude above every learned head (0.01–0.04), and its SWaT 0.90/0.54 is the strongest single family result in the ladder; equalizing every marginal costs it exactly where raw scale is the signal (Exathlon VUS-PR 0.59 against its siblings' 0.97). The NS-NN pair MADI3/MADI3_ENHANCED are solid generalists on compact, box-like modes (SMD best at 0.84) and fail together on drift (GHL VUS-PR 0.02); their 0.0004 AUC separation over 180 series is the ancillarity corollary as a measurement. CBCLA tracks that family where modes are compact (Exathlon 0.97 VUS-PR) and pays on plateau-heavy plants whose test regimes leave the training envelope (SWaT 0.35, PSM 0.37, GHL 0.43). The projection family reads cross-channel structure that axis-aligned boxes cannot: AE_LATENT is the best rung on OPPORTUNITY (0.65; body-worn sensor correlations) and strong on the ECG families, while CUBE_CERTIFY, the everything-composed rung, is never the specialist and rarely the victim (best of the ladder on MSL, 0.67): the robustness-for-peak trade of composing all seven blocks. Of the temporal lenses, DELAY_LATENT owns waveform and rhythm families (Daphnet 0.94, MITDB 0.77, SVDB 0.75, Exathlon 0.99) and is the ladder's closest per-series rival to xLSTMAD; RESID_LATENT is the most polarized rung: where a forecaster is learnable and the fault is an innovation it wins outright (GECCO 0.93/0.30 with every other rung $\leq$ 0.79/0.08; SMAP best at 0.78/0.33), and where the anomaly is itself a drift, the forecaster tracks it into invisibility (GHL 0.40, OPPORTUNITY 0.40, Genesis 0.24), the blind spot its CUSUM arm only partly repairs. Two humbling constants close the ledger: TAO's in-range point events hold every rung near chance

(0.50–0.58 AUC), and the 13 series carrying mixed point-and-sequence labels are hard for everything, our rungs and xLSTMAD alike (≈ 0.60).

Reading Figure 14. Rows are dataset families (one series each), identically ordered in both panels by mean AUC-ROC; columns are the ten rungs in ladder order. The two panels are the two currencies of Section 6.3: AUC-ROC has chance level 0.5 everywhere, while VUS-PR's chance level is the anomaly rate (0.005–0.02), so the same cell can be simultaneously strong in ranking and modest in delivery (Genesis: AUC 0.84, VUS-PR 0.04 at an anomaly rate of 0.3%). The family structure matches the theory's predictions: certified boxes dominate where modes are box-like (SMD, LTDB, Exathlon); the lens rungs carry drift-dominated families (GHL: CANONICAL 0.78 and LATENT_CANONICAL 0.79 after accumulator and setpoint routing, against ≤ 0.55 for every method before routing); and the hardest families for every rung (MSL, GHL in delivery terms) are exactly those whose anomalies are quiet, inside every marginal box, where pointwise information is provably thin.

I AIRCRAFT FLIGHT-TEST EXECUTION DETAILS

This appendix documents the campaign behind Section 6.4, Table 2, and Figure 15. The data derive from Sipple et al. (2026): real 1 Hz telemetry from the public NASA DASHlink Tail 6X fleet of regional commuter jets, with failure mechanisms seeded as deterministic perturbations of temporal windows inside otherwise real held-out flights.

Corpus and labels. Training positives are a consolidated corpus of unlabeled nominal operation from the fleet's 2001–2002 window: 336 flights, 2.02M rows, reduced to 1,687,706 rows (83.4% retention) by an asymmetric percentile glitch filter (drop below the 1.0th and above the 99.9th percentile per physical channel), which in the paper's terms tightens $\mathrm{supp}(P)$ toward Assumption (S). Test flights come from the disjoint 2003–2004 window: 630 candidate files read at 1 Hz under a 500k-point cap yield 75 flights: 12 nominal and 63 carrying one seeded mechanism each (12 compressor stall, 12 heat-exchanger fouling, 11 fan blade out, 11 sensor hunting, 11 HPT coolant loss, 6 progressive vibration), each with a per-sample label written by the failure simulator (we score $y = 1 - \texttt{class_label}$) and a milder per-flight glitch filter (0.5th percentile, lower side only). The seeded windows differ enormously in extent, so the pooled anomaly rate, VUS-PR's chance level, spans 0.003 (compressor stall) and 0.022 (sensor hunting) through 0.289 (vibration), 0.492 (HPT), 0.571 (fan blade out), and 0.701 (fouling); this is why Table 2 forbids comparing VUS-PR across mechanisms and Figure 15 prints π under each column.

Features and components. The physical state is 41 sensors: control surfaces, spool speeds (N_1, N_2), exhaust-gas temperatures, fuel flow and quantity, oil pressure and temperature, engine vibration, air data, body accelerations, and weight-on-wheels. The deployed network additionally consumes learned spectral context: six FFT coefficients each on the flap, two oil-pressure, and two vibration channels (30 columns) plus a weight-on-wheels categorical, giving 72 inputs routed to per-component subnetworks ($[128, 256]$ for the physical block, $[16, 16]$ per auxiliary component) under joint layers $[512, 64]$. The three NS-NN rungs (MADI3, MADI3_ENHANCED, CBCLA) consume all 72 inputs; the five certified rungs operate on the 41 physical sensors alone, because the certificates of Section 5 are statements about physical operating envelopes, not about learned representations. This asymmetry is deliberate and load-bearing for reading Figure 15: the two mechanisms the baseline wins are exactly the two whose evidence is spectral.

Dials. MADI3 trains with its production dials: sampling ratio $r_s = 30$, margin $\delta = 0.15$, dropout 0.2, learning rate 5×10^{-5}, 60 epochs (patience 20). The certified rungs reuse the design ledger: $K = 4$ operating modes (flight phases; k-means partitioner), $\gamma = \frac{1}{2}$, $\alpha = 0.05$, box margin 0.05, and an envelope budget of $r_s = 2$, the dimension-free shell economics of Theorem 4.2 against the baseline's 30. The quantile lens uses 257 knots per sensor, fitted on a dedicated half of the training rows with boxes certified on the disjoint remainder; C-BCLA runs 24 cooling rungs (3 ULA steps each, $\gamma_c = 0.85$) under exactly-spent budget $B_0 = 0.2$ with terminal mix $\zeta_{\max} = 0.25$ and a 20k-draw certificate at $\alpha = 0.05$; the projection rungs use $d_z = 8$ (encoder $[64, 32]$, conformal residual budget $\alpha_f = 0.01$, disjoint train/certify/conformal splits). Hardening matches Appendix H: log-space envelope normalizers, ρ_- floor 0.02, box trim 10^{-3}, small-pool guards, certificates that

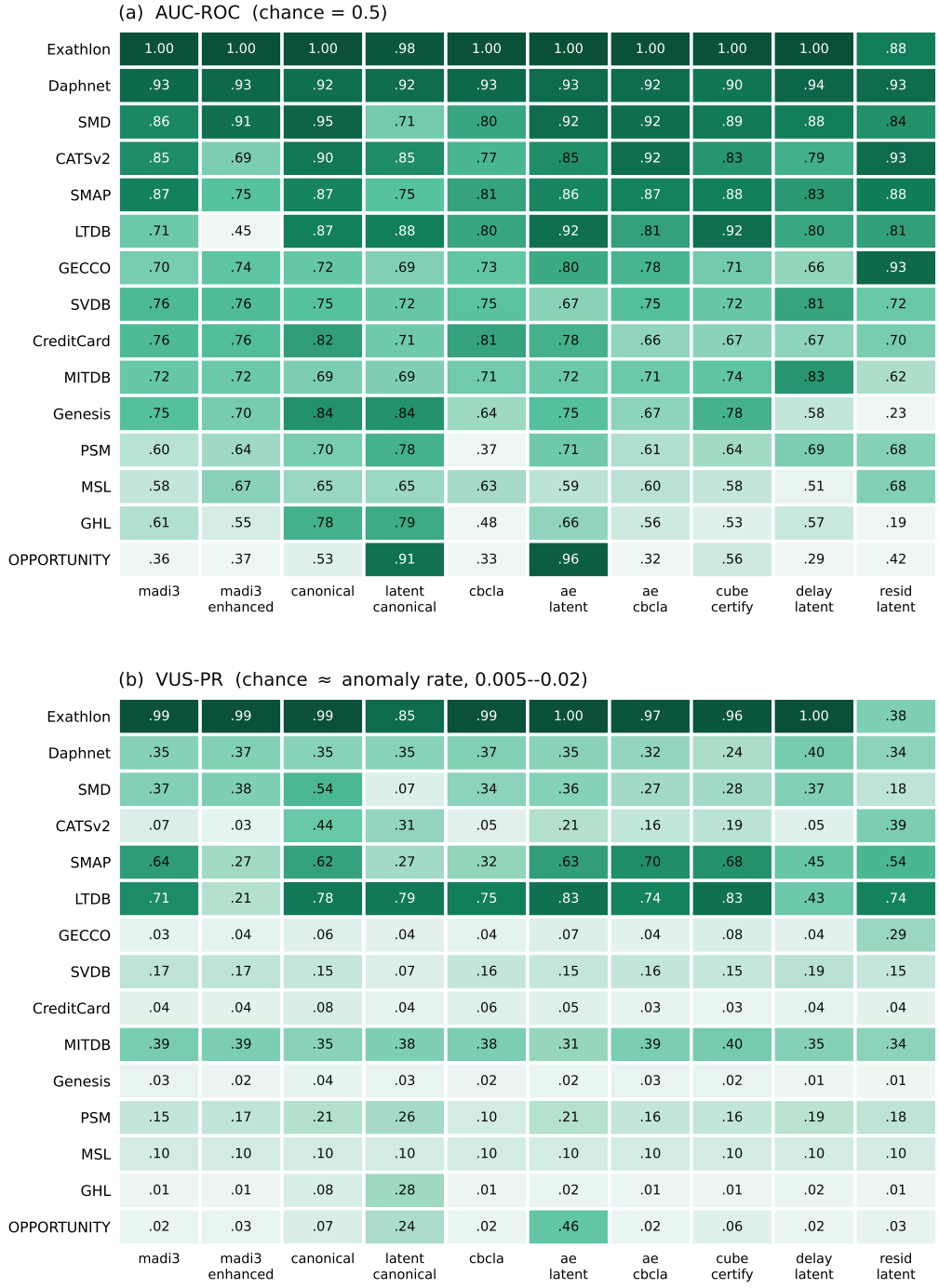


Figure 14: TSB-AD-M per-family results for the ten-rung ladder on the 15-series subset (one series from each of 15 of the 17 families; rows identically ordered in both panels). *(a)* AUC-ROC, chance $= 0.5$. *(b)* VUS-PR, whose chance level is the anomaly rate (0.005–0.02): the visibly lighter panel is the prevalence penalty of the precision currency, not a contradiction of panel (a); see Section 6.3.

degrade to CAUTION rather than abort. In this run the assembled rung's certificate *procedure* accepted its threshold without inflation ($\hat{\tau} = \tau = 2.41$) and the conformal residual set $\tau_f = 4.15$; per Remark E.1, the formal guarantee is conditional on hypotheses (box supports, flatness, i.i.d. rows, the phase partition, budgets) that flight telemetry does not verifiably satisfy, and rows within a flight are dependent, so the conformal budget α_f should be read at the flight level and the residual threshold

Rung	AUC-ROC	Rung	AUC-ROC
MADI3	$0.736\ [0.677, 0.791]$	CBCLA	$0.680\ [0.631, 0.728]$
MADI3_ENHANCED	$0.598\ [0.559, 0.637]$	AE_LATENT	$0.715\ [0.651, 0.778]$
CANONICAL	$0.759\ [0.705, 0.812]$	AE_CBCLA	$0.741\ [0.682, 0.798]$
LATENT_CANONICAL	$0.767\ [0.709, 0.823]$	CUBE_CERTIFY	$0.723\ [0.665, 0.780]$

Table 7: Flight-test campaign: mean per-flight AUC-ROC with 95% bootstrap CIs over the 63 failure flights ($B = 20{,}000$).

as a calibrated scale rather than a validated per-row false-alarm rate. Detectors train once on the consolidated corpus, in an order that defers each newly added rung to the end of the queue so every pre-existing detector consumes bit-identical random streams.

Scoring and metrics. Certified rungs score with the collar/union score; NS-NN rungs score $1-$class probability. Every score is delivered through a causal rolling maximum over one ACF period of the flight's leading sensor (the `find_length` estimator of Section 6.3, capped at 300 samples, fallback 125), matching the TSB-AD-M suite in both the delivery step and the window the VUS evaluator sweeps; the campaign's legacy evaluation (a centered 32-sample boxcar mean) remains available as a switch. Per-flight ROC-AUC is computed on the smoothed score; single-class flights are excluded from failure-type means. The orientation diagnostic (envelope negatives constructed in the inflated train box must out-rank held-out training rows) runs report-only here and is clean for all eight rungs: separation AUC 1.000 for six rungs, 0.999 and 0.981 for the other two, so no score was flipped.

Statistical uncertainty. Intervals are 95% percentile bootstrap CIs over the 63 failure flights ($B = 20{,}000$ flight resamples of mean per-flight AUC-ROC); Table 7 lists all eight rungs. Paired per-flight resampling gives LATENT_CANONICAL $-$ MADI3 $= +0.026$, CI $[-0.030, +0.087]$ (fraction of bootstrap mass above zero 0.80) and CANONICAL $-$ MADI3 $= +0.014$, CI $[-0.048, +0.076]$ (0.66). Both intervals cover zero: the overall lead of the certified lens rung over the deployed baseline is consistent across currencies but not resolved at $n = 63$ flights, and Section 6.4 states it accordingly. The mechanism-level contrasts that carry the section's reading are far larger than the overall gaps (fan blade out 0.989 versus 0.949; vibration 0.938–0.984 versus 0.605; sensor hunting 0.913 versus 0.61–0.80 the other way), with each contrast consistent in sign across nearly every flight of its mechanism.

Reading Figure 15. Rows are the eight rungs in Table 2's ladder order; columns are the six seeded mechanisms. The left panel (AUC) is comparable everywhere (chance 0.5); in the right panel each column has its own chance level π (printed beneath it), so cells are comparable down a column only: fouling's uniformly high VUS-PR (0.63–0.76) sits *at* its $\pi = 0.70$ chance and corroborates the left panel's verdict that no rung detects it, while compressor stall's modest-looking 0.32 is $\sim 100\times$ its $\pi = 0.003$. The two-regime structure of Section 6.4 is visible as vertical banding: certified rungs dark on fan blade out, vibration, and compressor stall; the spectral-input NS-NNs dark on sensor hunting and HPT coolant loss.

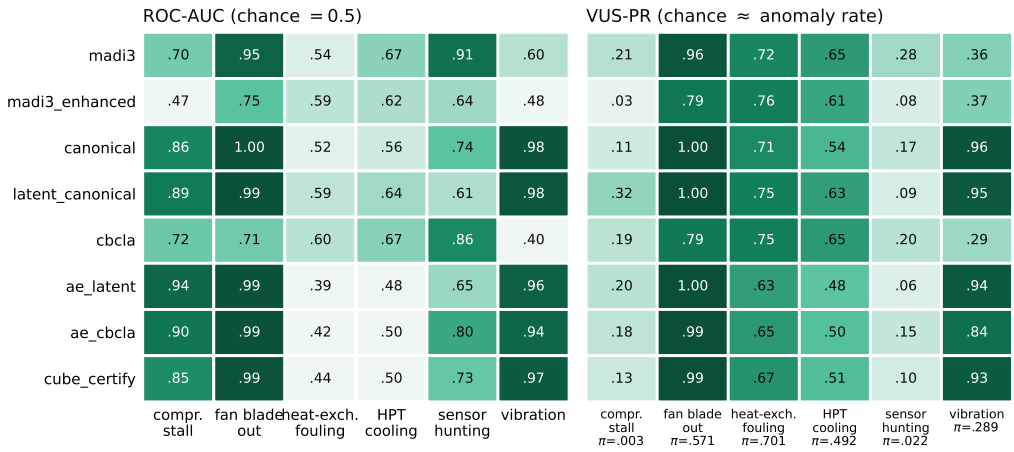


Figure 15: Flight-test campaign: mean per-flight ROC-AUC (left; chance 0.5) and VUS-PR (right; chance $= \pi$, printed per mechanism) by failure mechanism and detector. Mechanism labels abbreviate compressor stall and heat-exchanger fouling.