DETECTING TIME-SERIES ANOMALIES WITH SCORE MODELS ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and score-based generative models expose the gradient of the log-density, $s_p = \nabla \log p$, not the density itself. Can such score-only models detect anomalies in time series, and what exactly can they detect? First, the capability: a directly parameterized score network, scored by the window-averaged Hyvärinen statistic $J_{s_\theta} = \nabla \cdot s_\theta + \frac{1}{2}\|s_\theta\|^2$, reaches 0.588 ± 0.001 VUS-PR on the TSB-AD univariate benchmark, with *no statistically significant difference* from the strongest baseline PaAno (which stays ahead under a buffer-free control), at under a third of its measured runtime; on the multivariate suite it is mid-field. The divergence term is load-bearing and *significant*: $+0.023$ over the same network's score norm ($p = 0.025$). Second, the limits: we prove the population detection gap equals one half the Fisher divergence plus a signed Fisher-information correction, naming a blind spot of the upper-tail test that more data cannot remove: anomalies sharper than normal data invert the statistic whenever $I(q) - I(p) > F(q\|p)$, and we observe an inversion *consistent with* this class on a real sensor flat-line. Third, the mechanism: window aggregation estimates the statistic's population expectation (i.i.d. window model), and footprint-scale aggregation ($\approx L+w-1$) performs near the best tested scale in our synthetic length-group averages. Fourth, an equal-treatment audit: the same aggregation lifts five saved-score baselines by 0.02–0.23 VUS-PR and reorders the leaderboard (the lift surviving a buffer-free control), so part of the apparent ranking is output post-processing. In our matched energy-network comparison on the univariate suite, the two statistics show no statistically significant difference; the score-only statistic is useful when no energy readout is available.

1 INTRODUCTION

The strongest density estimators of the past five years do not estimate densities. Score-based and diffusion models (Song & Ermon, 2019; Ho et al., 2020) learn the score field $s_p = \nabla \log p$ by denoising score matching (Vincent, 2011) and expose no cheap $\log p$ (exact likelihoods need a probability-flow ODE solve per point, Song et al., 2021), so the classical recipe of flagging what is improbable (Ruff et al., 2021; Xu et al., 2018; Su et al., 2019) does not apply. Existing surrogates (the score norm, Mahmood et al., 2021; the diffusion-time posterior, Livernoche et al., 2024; a denoiser residual, Özer et al., 2026b) are each effective somewhere but none says what anomalous data does, on average, to the thresholded number. This paper asks: *can a score-only model detect time-series anomalies at the level of the best engineered systems, and what exactly can it detect?* The answer has four parts. The capability is an experiment: the Hyvärinen score $J_{s_\theta}(x) = \nabla \cdot s_\theta(x) + \frac{1}{2}\|s_\theta(x)\|^2$ (Hyvärinen, 2005) of a directly parameterized score network, computable from the score field alone with no energy in the model, run on TSB-AD (Liu & Paparrizos, 2024) under the strict protocol the benchmark's literature demands (Kim et al., 2022; Wu & Keogh, 2023; Paparrizos et al., 2022). The limits are a theorem: the population detection gap is *exactly* $\frac{1}{2}F(q \,\|\, p) + \frac{1}{2}(I(p) - I(q))$ (Corollary 1), a signed Fisher-information correction naming a blind spot of the upper-tail test that more data cannot remove: anomalies sharp enough that $I(q) - I(p) > F(q \,\|\, p)$ invert the statistic. Between them sits the mechanism (window aggregation estimates the score's expectation, and footprint-scale aggregation performs near the best tested scale in our synthetic length-group averages) and an equal-treatment audit showing part of the apparent leaderboard ranking is output post-processing. Our matched energy-network comparison finds no statistically significant difference on the univariate

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.

suite (Section 6); the windowed Hyvärinen statistic is a principled alternative when no energy readout is available.

Contributions.

- **Score-only detection competitive with the strongest baseline.** A directly parameterized score network scored by window-averaged J_{s_θ} reaches 0.588 ± 0.001 VUS-PR on TSB-AD-U, no statistically significant difference from PaAno (which leads under the buffer-free AUC-PR control), at under a third of PaAno's runtime; the divergence term adds a *significant* $+0.023$ over the score-norm statistic ($p = 0.025$, Section 6).
- **An exact, signed characterization of what score statistics detect.** The detection gap equals $\frac{1}{2}F(q\|p) + \frac{1}{2}(I(p) - I(q))$; upper-tail detection is guaranteed for $I(q) \leq I(p)$ and inverted when $I(q) - I(p) > F$, a blind spot consistent with the inversion we observe on a real flat-line (Corollary 1, Figure 5, appendix).
- **An empirically consistent design rule for output aggregation.** Detection follows a rise-then-dilute law; aggregating at the anomaly's score footprint ($\approx L+w-1$) is near-optimal on ground truth (within 0.026 AUC-PR of the dense-grid optimum for $L \geq 3$; the $L=1$ argmax is alignment-dominated); the Hutchinson probe count is immaterial after aggregation (Section 6).
- **An equal-treatment audit of TSB-AD.** Identical output aggregation lifts five saved-score baselines by 0.02–0.23 VUS-PR and reorders the leaderboard, the lift surviving a buffer-free pointwise control: part of the apparent ranking is output post-processing (Section 6).

2 Related Work

Score/energy models, OOD, and the Hyvärinen score. Score matching estimates an unnormalized density from $\nabla_x \log p$ without the partition function (Hyvärinen, 2005); DSM, which trains our networks, makes it tractable via the noise-smoothed density (Vincent, 2011; Song & Ermon, 2019). Detectors on such models read anomalies off the score norm (MSMA, Mahmood et al., 2021) or off an unnormalized model's energy/reconstruction (Zhai et al., 2016; Liu et al., 2020; Grathwohl et al., 2020) ("energy" $=$ unnormalized $-\log p$). MSMA is the sharpest comparison: our $\frac{1}{2}\|s\|^2$ ablation is its single-scale analogue, and we run an architecture-matched MSMA-style multiscale surrogate (single-Gaussian auxiliary) as an additional comparison (Section 6). The Hyvärinen score is a strictly proper local scoring rule (Parry et al., 2012), used for model comparison (Shao et al., 2019), and is related to kernelized-Stein tests (Liu et al., 2016; Chwialkowski et al., 2016); its closest relative is score-based CUSUM, governed by the same Fisher divergence in our identity (Wu et al., 2023b; Chen et al., 2025), but for sequential change detection with distinct pre/post distributions rather than our unsupervised windowed detection with a single normal model; we add an exact detection-gap identity, concentration, and conformal validity (a per-window KSD test costs $O(m^2)$ vs. our $O(m)$). The statistic also connects to likelihood-free OOD/typicality detection (Nalisnick et al., 2019; Huang et al., 2021; Ren et al., 2019b; Serrà et al., 2020). We found no prior work developing the Hyvärinen score as an unsupervised anomaly scoring rule with a window-level guarantee.

Time-series/diffusion anomaly detection, benchmarks, and aggregation. A parallel line scores anomalies from diffusion/denoising models: DTE (Livernoche et al., 2024), JuRe (Özer et al., 2026b), and reconstruction-based approaches (Wyatt et al., 2022; Wang et al., 2023; Chen et al., 2023) provide alternative anomaly readouts rather than the full Hyvärinen statistic; our matched score-norm ablation separately isolates the divergence term's contribution. We evaluate on TSB-AD (Liu & Paparrizos, 2024), which recommends the threshold-free VUS-PR headline (Paparrizos et al., 2022) resting on critiques that point-adjustment and best-of-N selection manufacture rankings (Kim et al., 2022; Wu & Keogh, 2023; Sarfraz et al., 2024; Lyu, 2026). Our baselines span deep reconstruction/forecasting (Audibert et al., 2020; Tuli et al., 2022; Xu et al., 2022; Wu et al., 2023a; Su et al., 2019; Xu et al., 2018; Malhotra et al., 2015; Munir et al., 2019; Sakurada & Yairi, 2014; Xu et al., 2024; Faber et al., 2025), recent leaderboard methods (Park & Kang, 2026; Özer et al., 2026a; Lan et al., 2026; Özer et al., 2026c; Yeh, 2026), and the classical family (Liu et al., 2008; Breunig et al., 2000; Schölkopf et al., 2001; Yeh et al., 2016; Boniol et al., 2021; Paparrizos & Gravano, 2015; Shyu et al., 2003; Goldstein & Dengel, 2012; Li et al., 2020; Hariri et al., 2021; He et al., 2003; Ramaswamy et al., 2000; Rousseeuw & Van Driessen, 1999; Ren et al., 2019a;

Candès et al., 2011); PaAno (U) and AxonAD (M) are the strongest recent references. Other recent detectors (Zhou et al., 2025; Wu et al., 2025; Shentu et al., 2025; Yang et al., 2023; Kim et al., 2024) are not re-run. Pretrained time-series models (Goswami et al., 2024; Ansari et al., 2024; Das et al., 2024; Rasul et al., 2023; Ekambaram et al., 2026; Zhou et al., 2023; Dutta et al., 2025) enter as cite-only rows (Appendix I). Aggregating before alarming is long-established (EWMA and smoothed errors, Roberts, 1959; Hundman et al., 2018; Ren et al., 2019a; subsequence detectors, Yeh et al., 2016; Boniol et al., 2021; matched filtering, Turin, 1960; window-size selection, Ermshaus et al., 2023; collective anomalies, Chandola et al., 2009; batch typicality tests, Nalisnick et al., 2019); we formalize it as an unbiased estimator of the exact signed detection-gap functional (Corollary 1). Conformal prediction gives distribution-free control under exchangeability (Vovk et al., 2005; Bates et al., 2023; Laxhammar & Falkman, 2015); we report that exchangeability fails between a series prefix and its remainder, so naive calibration over-alarms under drift, consistent with heterogeneous per-series additive gaps (Barber et al., 2023) (Section 6).

3 METHOD

3.1 BACKGROUND AND PROBLEM STATEMENT

The *score* of a density p is $s_p := \nabla \log p$, invariant to the normalizing constant. An *energy-based model* sets $p \propto \exp(-E)$, so the *energy* $E = -\log p + \text{const}$ is the unnormalized negative log-density and $s_p = -\nabla E$ (LeCun et al., 2006); this is what "energy" means throughout. *Score matching* (Hyvärinen, 2005) fits s_θ to s_p without Z: under Assumption 1's integration-by-parts conditions applied to s_θ, $\mathbb{E}_p[J_{s_\theta}] = \frac{1}{2}\mathbb{E}_p\|s_\theta - s_p\|^2 - \frac{1}{2}I(p)$, so minimizing $\mathbb{E}_{X\sim p}[J_{s_\theta}(X)]$ minimizes the score-matching loss, with $J_{s_\theta} = \nabla\cdot s_\theta + \frac{1}{2}\|s_\theta\|^2$, the *Hyvärinen score*, which we repurpose as an anomaly statistic. *Denoising score matching* (DSM) (Vincent, 2011) fits the score of the noise-smoothed p_σ in closed form (Vincent, 2011; Song & Ermon, 2019), the regime of score/diffusion models, which expose a score but no closed-form $\log p$ (probability-flow likelihoods exist at the cost of an ODE solve per point; Song et al., 2021).

We observe a C-channel series and slide a length-w window, flattened to $x \in \mathbb{R}^d$ ($d = wC$) and z-scored using train-prefix statistics only. Write p for the density of normal windows (positive, C^2) and q for a possibly anomalous source.

Definition 1 (Hyvärinen score)**.** For a positive C^2 density p, the (pointwise) *Hyvärinen score* is

$$J_p(x) \;:=\; \nabla^2 \log p(x) + \tfrac{1}{2}\left\|\nabla \log p(x)\right\|^2 \;=\; \nabla\cdot s_p(x) + \tfrac{1}{2}\left\|s_p(x)\right\|^2, \tag{1}$$

computable from s_p alone and free of Z (Hyvärinen, 2005) ($\nabla^2 = \nabla\cdot\nabla$ denotes the Laplacian). The Fisher divergence is $F(q\,\|\,p) := \mathbb{E}_{x\sim q}\|\nabla \log q(x) - \nabla \log p(x)\|^2 \geq 0$, zero iff $q = p$ (under regularity).

We use J_p as a *signed* deviation score: a window from q shifts its mean by the signed detection gap of Corollary 1, so upper-tail use is justified only for a positive gap (a concentrated anomaly can score *below* normal data). Section 4 makes this precise for *window averages*, where the finite-sample statement applies under its i.i.d. model (the centered average uses future scores: the headline evaluation is *offline*, matching whole-series scoring); a *window mode* (headline, Theorems 1, 2) and a secondary *point mode* (heuristic) follow.

3.2 ESTIMATOR

We never form p; we fit a score network to normal windows and read J_p off it, using two parameterizations. The *headline* is a directly parameterized score network $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ (the diffusion-model architecture class) for which no energy or $-\log p$ exists anywhere in the model; the plug-in bridge (Proposition 1) covers it since it assumes no conservativity. The second is an *energy-parameterized* network ($s_\theta = -\nabla E_\theta$), used for the energy-versus-J_p comparison of Section 6 and several mechanism experiments (coercive-anchor construction in Appendix B.1).

Denoising score matching. We train s_θ by DSM (Vincent, 2011) over a geometric noise ladder $\sigma \in [0.05, 1.0]$: for $x_\sigma = x + \sigma\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$, the σ^2-weighted loss is

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{x,\sigma,\varepsilon}\; \tfrac{1}{2}\left\|\, \sigma\, s_\theta(x + \sigma\varepsilon, \sigma) + \varepsilon \,\right\|^2, \tag{2}$$

whose minimizer satisfies $s_\theta(\cdot, \sigma) \approx \nabla \log p_\sigma$ (Vincent, 2011; Song & Ermon, 2019). At test time we evaluate at a single scale σ_{eval} selected on a disjoint tuning split (Section 5).

Divergence estimate. The squared-norm term of equation 1 is free once $s_\theta(x)$ is evaluated; the divergence $\nabla \cdot s_\theta$ is computed exactly or by Hutchinson's estimator with K Rademacher probes (Hutchinson, 1990) (cost/complexity Proposition 3, Appendix A.5). We use the exact trace when $d \leq 64$ and $K = 4$ probes otherwise; window aggregation reduces the probe noise, and benchmark performance was insensitive to the tested probe counts (Section 6).

Procedure. In one line: **train** s_θ on normal data with the DSM loss equation 2; **calibrate** the conformal threshold as the $\lceil (1-\alpha)(n+1) \rceil$-th order statistic of the window means $T(W) = \frac{1}{|W|} \sum_{x \in W} \hat{J}(x)$; **score** each test window by $T(W)$ and its conformal p-value (window mode, headline; $\hat{J}(x)$ a secondary point mode).

4 Theory

All results are stated for population densities p, q; the estimator of Section 3.2 plugs in for s_p. We assume once the standard regularity conditions (p, q positive and C^2, finite Fisher-type moments, vanishing boundary flux; full statement Assumption 1, Appendix A.1), which justify $\mathbb{E}_q[J_q] = -\frac{1}{2} I(q)$ and $\mathbb{E}_p[J_p] = -\frac{1}{2} I(p)$; full proofs are in the appendix.

4.1 Exact separation

Lemma 1 (Integration by parts)**.** *Under the regularity conditions above,* $\mathbb{E}_q[s_q^\top s_p] = -\mathbb{E}_q[\nabla \cdot s_p]$.

Proof sketch. Integrate by parts using the vanishing boundary flux of $q\, s_p$; full proof in Appendix A.

Theorem 1 (Classical score-matching identity (Hyvärinen, 2005))**.** *Under the regularity conditions above,*

$$\boxed{\mathbb{E}_q[J_p] - \mathbb{E}_q[J_q] \;=\; \tfrac{1}{2} F(q \,\|\, p) \;\geq\; 0} \tag{3}$$

with equality iff $q = p$.

Proof sketch. Expand $F(q\|p)$, substitute Lemma 1, and subtract the $p = q$ specialization; full proof in Appendix A. The identity is exact (constant $\frac{1}{2}$, genuine divergence); we verify it numerically to relative error $\leq 2.3 \times 10^{-3}$ (appendix).

From the identity to a detection gap. Detection compares the same statistic on normal and anomalous data. Writing $I(r) := \mathbb{E}_r \|\nabla \log r\|^2$ for Fisher information, Theorem 1 with $\mathbb{E}_p[J_p] = -\frac{1}{2} I(p)$ and $\mathbb{E}_q[J_q] = -\frac{1}{2} I(q)$ gives what detection actually rests on:

Corollary 1 (Detection gap)**.** *Under the regularity conditions above,*

$$\mathbb{E}_q[J_p] - \mathbb{E}_p[J_p] \;=\; \tfrac{1}{2} F(q \,\|\, p) \;+\; \tfrac{1}{2}\big(I(p) - I(q)\big). \tag{4}$$

The signed information term is a blind spot of upper-tail *mean-based* separation, persisting in the population limit (a mean shift, not a per-window statement; finite-window conclusions remain conditional on concentration; Figure 2, appendix). Writing Δ for the gap equation 4, if $I(q) \leq I(p)$ then $\Delta > 0$ for every $q \neq p$, so upper-tail thresholding separates this class (sufficient, not necessary); but anomalies *sharper* than normal data can make the gap zero or negative, and upper-tail detection is *inverted precisely when* $I(q) - I(p) > F(q\|p)$, a class consistent with observed inversions on real low-variance flat-lines (exact $I(q) - I(p)$ versus $F(q\|p)$ balance not estimated; for $p = \mathcal{N}(0,1)$, $q = \mathcal{N}(0, s^2)$ gives $\Delta = (s^2 - 1)/2$, negative iff $s^2 < 1$). A two-sided test is the natural route to negative-gap alternatives (an exploratory variant that forfeits conformal validity and does not improve accuracy on TSB-AD, Sections 4.4, 6). We observe an inversion *consistent with* this named class on pattern-loss synthetic anomalies and real flat-lines (Section 6), and equation 3 governs *expectations* under q, not points, so window-level detection is the headline and point mode a heuristic.

Relation to the likelihood and Gaussian check. The likelihood has the same signed structure ($(F, I) \to (\mathrm{KL}, H)$: $\mathbb{E}_q[-\log p] - \mathbb{E}_p[-\log p] = \mathrm{KL}(q\|p) + H(q) - H(p)$), so J_p is not a repackaged likelihood (high rank correlation but structured disagreement, Table 14); for Gaussian p it is an affine Mahalanobis-type distance, ranking identically to $-\log p_\sigma$ for isotropic covariance at any noise scale, and for general Σ only in the large-noise limit (Proposition 2, Corollary 3, appendix).

4.2 Window concentration

Averaging J_p over m points estimates the controlled expectation with error $\sim 1/\sqrt{m}$; a competing effect is that an anomaly shorter than the window occupies a fraction $\rho \approx (L+w-1)/m$ (its score footprint), so normal points dilute the average by ρ. Under an i.i.d. mixture these combine into a signal-to-noise proxy predicting a rise-then-dilute shape. With J_p sub-exponential under q with parameters (ν, b) (Appendix A.5), a Bernstein bound gives $\Pr[|\frac{1}{m}\sum_i J_p(x_i) - \mathbb{E}_q J_p| > t] \le 2\exp(-c\,m\min(t^2/\nu^2, t/b))$, so the window size needed to separate the anomalous mean from the normal value at confidence $1-\delta$ scales as follows.

Theorem 2 (Window concentration). *Let an anomalous window $x_1, \dots, x_m$ be i.i.d. from q and an independent normal window $x'_1, \dots, x'_m$ be i.i.d. from p, with J_p sub-exponential under both laws with parameters (ν, b) (the maxima over the two), and write $\Delta := \mathbb{E}_q[J_p] - \mathbb{E}_p[J_p] = \frac{1}{2}F(q\|p) + \frac{1}{2}(I(p) - I(q))$ (Corollary 1). Whenever $\Delta \neq 0$, separation (both window means strictly on their own population's side of the two-mean midpoint; union bound) at confidence $1-\delta$ is guaranteed once the window size exceeds*

$$m^\star \gtrsim \max\left\{\frac{\nu^2}{\Delta^2}, \frac{b}{|\Delta|}\right\} \log\frac{1}{\delta} \tag{5}$$

up to absolute constants (full tail bound Appendix A.5).

Equation 5 is sufficient, not necessary: the window size scales as ν^2/Δ^2 in the sub-Gaussian Bernstein regime and as $b/|\Delta|$ in the sub-exponential regime (the Fisher divergence is only the nonnegative component of Δ). The near-anomaly-length peak comes not from this Bernstein bound (no anomaly length) but from the mixture: if only a fraction ρ of a length-m window is anomalous, linearity gives effective separation $\rho\,\Delta$, so enlarging m past the anomaly length lowers ρ and dilutes the signal: detection rises then degrades, the variance-reduction-versus-dilution trade-off Section 6 confirms on real series and a synthetic benchmark with known lengths (dependence/β-mixing extension in Remark 1, appendix).

4.3 From the exact score to the learned score

The detector uses a learned field s_θ trained by DSM against p_σ; the identity and gap corollary apply verbatim with p_σ in place of p, leaving estimation error $e := s_\theta - s_{p_\sigma}$. Here $\Delta := \mathbb{E}_q[J_{p_\sigma}] - \mathbb{E}_{p_\sigma}[J_{p_\sigma}]$ and $\Delta_{s_\theta} := \mathbb{E}_q[J_{s_\theta}] - \mathbb{E}_{p_\sigma}[J_{s_\theta}]$ are smoothed-reference gaps (the deployed clean-p gap is equation 7).

Proposition 1 (Estimation error of the gap). *Under L^2-integrability and zero-boundary-flux conditions on e for $r \in \{q, p_\sigma\}$ (Appendix A.7), the detection gap of the plug-in statistic J_{s_θ} deviates from Δ by at most*

$$\left|\Delta_{s_\theta} - \Delta\right| \le \sqrt{F(q\,\|\,p_\sigma)}\,\|e\|_{L^2(q)} + \tfrac{1}{2}\Big(\|e\|^2_{L^2(q)} + \|e\|^2_{L^2(p_\sigma)}\Big). \tag{6}$$

The deviation is first order in the L^2 score error (constant $\sqrt{F}$) and second order otherwise; DSM controls $\|e\|_{L^2(p_\sigma)}$. Crucially equation 6 assumes *no conservativity*, hence covers the directly parameterized network, for which s_θ need not be a gradient.

The operational gap: clean data against a smoothed reference. The deployed detector scores *clean* normal windows (density p) with a statistic built from s_{p_σ}, so the gap it operates on is $\mathbb{E}_q[J_{p_\sigma}] - \mathbb{E}_p[J_{p_\sigma}]$; applying Theorem 1 under both q and p,

$$\mathbb{E}_q[J_{p_\sigma}] - \mathbb{E}_p[J_{p_\sigma}] = \tfrac{1}{2}\Big[F(q\,\|\,p_\sigma) - F(p\,\|\,p_\sigma)\Big] + \tfrac{1}{2}\big(I(p) - I(q)\big). \tag{7}$$

The subtracted $F(p \,\|\, p_\sigma) \geq 0$ is the *normal-reference mismatch correction*, vanishing as $\sigma \to 0$; smoothing also changes the anomaly term ($F(q\|p) \to F(q\|p_\sigma)$), so the subtracted normal-reference term alone does not capture the full smoothing effect (equation 7 itself is the complete operational gap). Tuning-split selection targets empirical detection performance measured by VUS-PR. The finite-sample statements apply to equation 7 under their stated assumptions (independent draws, sub-exponential J_{s_θ}); the overlapping-window pipeline adds dependence they do not cover (Appendix A.5).

Corollary 2 (End-to-end operational gap of the deployed statistic)**.** *Let $e := s_\theta - s_{p_\sigma}$ and let the integrability and zero-boundary-flux conditions of Proposition 1 hold for $r \in \{q, p\}$. Then the gap the deployed detector operates on satisfies*

$$\left|\big(\mathbb{E}_q[J_{s_\theta}] - \mathbb{E}_p[J_{s_\theta}]\big) - \Delta_{\mathrm{op}}\right| \;\leq\; \sqrt{F(q \,\|\, p_\sigma)}\,\|e\|_{L^2(q)} + \sqrt{F(p \,\|\, p_\sigma)}\,\|e\|_{L^2(p)} + \tfrac{1}{2}\|e\|^2_{L^2(q)} + \tfrac{1}{2}\|e\|^2_{L^2(p)}, \tag{8}$$

where Δ_{op} is the population operational gap of equation 7. The bound controls the effect of score-estimation error relative to this smoothed-score reference; the effect of smoothing itself is described by equation 7. Writing $\Delta_\theta := \mathbb{E}_q[J_{s_\theta}] - \mathbb{E}_p[J_{s_\theta}]$, Theorem 2 applies to J_{s_θ} with gap $\Delta_\theta \neq 0$, provided its sampling and sub-exponential assumptions hold for the learned statistic (assumed, not inherited).

Proof sketch. Apply the expectation identity of Appendix A.7 under q and p, subtract, and bound cross terms by Cauchy–Schwarz; see Appendix A.8. The bound is an *interpretation*, not an operational guarantee: it is conditional on score error under q, which normal-data DSM does not control, and extension to dependent windows requires additional assumptions (Appendix A.5).

4.4 Conformal calibration

We calibrate the window statistic T by split conformal prediction on held-out normal windows and alarm when the conformal p-value $\leq \alpha$. The headline uses the one-sided upper-tail statistic; because the detection gap can be negative, we also report an exploratory two-sided statistic $|T - \mathrm{med}(T_{\mathrm{cal}})|$.

Theorem 3 (Conformal validity)**.** *If, conditional on the fitted model and preprocessing (score network, normalization, and window construction fixed without using calibration or test observations differently), the calibration scores and a test score are exchangeable, then the split-conformal p-value π satisfies $\Pr[\pi \leq \alpha] \leq \alpha$ for every $\alpha \in (0, 1)$, giving finite-sample, distribution-free false-alarm control (Vovk et al., 2005; Bates et al., 2023), regardless of how well s_θ fits p.*

Two caveats limit this in practice: exchangeability between a training prefix and the series remainder *fails empirically* under drift, so naive split conformal over-alarms (Section 6); and the exploratory two-sided statistic forfeits the finite-sample guarantee because it estimates its median center from the same calibration windows it ranks against (Appendix A.6); two-sided scoring itself is not the obstacle (a center fitted on a separate training split would restore validity under exchangeability), so the guarantee applies cleanly only to our one-sided rank statistic.

5 Experimental setup

We evaluate on TSB-AD (Liu & Paparrizos, 2024), a benchmark built to correct known evaluation problems, with a univariate suite (TSB-AD-U, 350 series) and a multivariate suite (TSB-AD-M, 180 series) over twenty-plus sources. Our headline statistic is the window-averaged Hyvärinen score $J_{s_\theta} = \nabla \cdot s_\theta + \frac{1}{2}\|s_\theta\|^2$ of a *directly parameterized* score net ("J_{s_θ}, direct score net"), for which no energy exists anywhere in the model. To isolate score-only capability from any likelihood advantage we also train an energy-parameterized network and read two statistics off its shared weights, its window-averaged energy ("DSM energy") and its Hyvärinen score J_p (K=4), so the energy-versus-J_p comparison (Section 6) runs with no confound. The protocol below is fixed in advance, with one disclosed exception (the σ researcher degree of freedom, end of this section); full estimator/architecture/training details are in Appendix B.1.

- **Benchmark and split.** Semi-supervised: detectors fit only on the train prefix, no labels reach any detector, z-scoring/calibration use train-prefix statistics only (the label-free period estimate uses

Table 1: TSB-AD leaderboards, compact form (VUS-PR, mean over eval series; mean ± std over seeds $\{2024, 2025, 2026\}$; single-seed rows carry no std; sorted by VUS-PR). Ours: J_{s_θ} and J_p at σ=0.35, DSM energy at σ=0.2 (each statistic's tuning-split optimum). Cite-only rows, taken from the bundled TSB-AD CSVs (Liu & Paparrizos, 2026) under the same metric/split, are marked † and carry no paired test. Each panel includes its RANDOM control and coverage n. Expanded baseline leaderboards with the AUC-PR, VUS-ROC, and AUC-ROC secondary metrics are Tables 6 and 8; the direct-net headline row (0.588 U / 0.427 M) and its component ablations (indented U sub-rows: score-norm only, divergence only, pointwise) are also tabulated in Table 18.

TSB-AD-U (n=350)		**TSB-AD-M** (n=180)	
Method	VUS-PR	Method	VUS-PR
J_{s_θ} direct net (ours)	$\mathbf{0.588 \pm 0.001}$	AxonAD (162 ser., seed 2024)	**0.494**
score-norm only	0.565 ± 0.001		
divergence only	0.456 ± 0.011		
pointwise J_{s_θ}	0.506 ± 0.001		
Time-RCD (MAFT)†	0.586	DSM energy (ours)	0.458 ± 0.001
PaAno	0.584 ± 0.000	J_p (ours)	0.436 ± 0.005
DSM energy (ours)	0.577 ± 0.004	StreamVAE	0.428 ± 0.001
J_p (ours)	0.561 ± 0.005	J_{s_θ} direct net (ours)	0.427 ± 0.002
Sub-KNN	0.529	JuRe	0.423 ± 0.003
DTE	0.463 ± 0.004	PaAno	0.405 ± 0.000
RANDOM	0.086 ± 0.000	RANDOM	0.096 ± 0.000

the full series, identical for all methods, Appendix B), the whole series is scored. We inherit the benchmark's caveats unchanged: 35/350 (U) and 20/180 (M) train prefixes contain anomalies.

- **Metric.** Headline VUS-PR (the TSB-AD study's most robust metric), AUC-PR secondary; we do not rank on ROC (a uniform-random scorer reaches VUS-ROC 0.627/0.579 yet VUS-PR near the anomaly rate). The main leaderboards and audit include a RANDOM row.
- **Seeds.** Our statistics and the re-run neural baselines on the main leaderboards use three pre-registered seeds $\{2024, 2025, 2026\}$ (mean ± std, no best-of-N). The equal-treatment aggregation runs and the AxonAD multivariate re-run use seed 2024 only (an interrupted seed-2025 AxonAD-M re-run, 120/180 series, is reported in the supplementary material but not pooled).
- **Hyperparameter selection.** σ_{eval} and all detector hyperparameters are selected on the disjoint 48-series TSB-AD-U tuning split (tuning ∩ eval = ∅, verified) and reused unchanged on M (no M tuning split exists): J_p and J_{s_θ} use σ=0.35, the windowed energy σ=0.2, each statistic's tuning optimum (one disclosed researcher degree of freedom below).
- **Baselines (run vs. cite).** We re-run, under the identical protocol with the RANDOM control: PaAno, Time-RCD, DTE, StreamVAE, AxonAD, JuRe, MMPAD, USAD, TimesNet, TranAD, OmniAnomaly, AnomalyTransformer, FITS, CNN, LSTMAD, AutoEncoder, plus the bundled classical/subsequence detectors (catalog, settings, checks, failures, and one scaler-bug patch: Appendix I). Where a method cannot run under our protocol we cite its bundled TSB-AD number under the same metric/split (Liu & Paparrizos, 2026) and mark it cite-only: Time-RCD (MAFT) 0.586, TSPulse-FT 0.552, CHARM 0.501, xLSTMAD 0.398.

6 RESULTS

Main leaderboards. On both suites our statistics sit among the strongest methods (Table 1; expanded baseline leaderboards Tables 6, 8, direct-net rows Table 18). On TSB-AD-U the directly parameterized score net's J_{s_θ} has the highest observed mean at 0.588 ± 0.001 VUS-PR among our re-runs, with *no statistically significant difference* from PaAno (0.584): the paired difference over the 350 U series is +0.004 with a per-series 95% bootstrap CI of $[-0.032, +0.041]$ (Wilcoxon $p = 0.64$; source-clustered CI $[-0.109, +0.092]$). The nearest published entrant, cite-only Time-RCD (MAFT, 0.586), lies 0.002 below but admits no paired test. The energy-net readouts sit just below (windowed DSM energy 0.577 ± 0.004; J_p 0.561 ± 0.005). On TSB-AD-M the direct-net J_{s_θ} is *mid-field* (0.427 ± 0.002); the capability result there is carried by the energy-net windowed energy (0.458 ± 0.001, above every full-coverage baseline); AxonAD (published #1) reaches 0.494 on the 162/180 series it completes, where our windowed energy scores 0.505. RANDOM: 0.086 (U), 0.096 (M).

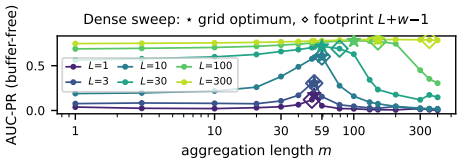


Figure 1: Dense footprint sweep on the synthetic benchmark (buffer-free AUC-PR, windowed J_p, 480 configurations, 17-point m grid): star = grid optimum, open diamond = footprint $L+w-1$. The optimum equals the footprint at $L=3, 10$ (at $L=1$ the argmax is alignment-dominated), and sits below it on a near-flat curve for $L \geq 30$; real-data sweep: Table 15.

Table 2: What aggregation is worth. VUS-PR on TSB-AD-U (energy-parameterized network, 3 seeds, $\sigma = 0.35$) for each statistic scored pointwise and then aggregated over $m =$ `slidingWindow` points; J_p: 350 series; energy: the 349 that completed; component rows: pointwise 350, aggregated 349. Gains are computed before rounding.

Statistic (same trained network)	pointwise	aggregated	gain
energy $= -\log p$ (unnormalized)	0.518 ± 0.002	$\mathbf{0.570 \pm 0.002}$	$+0.052$
J_p (score only)	0.473 ± 0.008	0.561 ± 0.005	$+0.087$
divergence term $\nabla \cdot s$ only	0.420 ± 0.005	0.482 ± 0.015	$+0.063$
score norm $\frac{1}{2}\|s\|^2$ only	0.368 ± 0.008	0.439 ± 0.012	$+0.071$
RANDOM	0.086 ± 0.000		–

Score-only detection. A model that exposes only a score is competitive with the strongest engineered baseline: the window-aggregated J_{s_θ} of a directly parameterized score net (no energy in the model) shows no significant difference from PaAno on TSB-AD-U (paired TOST-equivalent within ± 0.05, p=0.007, but not ± 0.03, p=0.085; post-hoc margins); PaAno stays ahead under buffer-free AUC-PR (0.552, Table 6, vs. 0.521, Table 3); the direct net runs at 3.0 s per series (fit + score, divergence included) against PaAno's 10.1 s under the identical protocol on the same hardware (Appendix I). Table 1 decomposes the statistic (full 0.588; score-norm alone 0.565; divergence alone 0.456; pointwise 0.506): the divergence term adds $+0.023$ (CI $[+0.007, +0.039]$, $p = 0.025$, Table 5).

Coverage, seed, and researcher-degree-of-freedom disclosures. The direct-score headline and most re-run baselines use three seeds and full coverage (per-row n/seeds: Tables 6–8). Exceptions: DTE 322/350 in the audit only (leaderboard 350/350); StreamVAE-U 318/350 (all-seed cohort); energy-U 349/350 at σ=0.05 and 0.35 (σ=0.2: 350/350); LSTMAD 274 (U) and 147/180 (M); MMPAD 10/350 (U) and 139/180 (M), single seed; AxonAD-M 162/180, AxonAD-U single-seed; Time-RCD U-only single-seed (failures: Appendix I); audit + AxonAD re-runs seed 2024. Researcher degree of freedom: the σ sweep was prompted by an eval-set observation at σ=0.05; the full eval-set sweep is ablation-only; the headline σ=0.35 is fixed on the tuning split.

When an energy is available. For the energy-parameterized network at matched σ=0.35, windowed energy exceeds windowed J_p by 0.009 VUS-PR on U, *not statistically significant* (Wilcoxon $p = 0.051$, $n = 349$; paired bootstrap CI for energy$-J_p$: $[-0.017, +0.034]$); on M the difference is 0.018 favoring energy. On the tuning split J_p peaks at σ=0.35 (gap to energy -0.007, Table 11). These results do not establish an advantage for replacing an available energy readout with J_p. The two statistics converge in performance, not in identity (median per-series Spearman ≈ 0.90, structured residual disagreement, Table 14); on the tuning split, multiscale J_p trails the best single σ (0.489 vs. 0.512).

The mechanism: footprint-matched aggregation. Aggregation turns pointwise Hyvärinen scores into an estimator of the expectation appearing in the detection-gap identity; footprint-scale windows perform near the measured optimum in our synthetic length-group averages. The moving average over m points improves every score-based statistic by 0.05–0.09 VUS-PR (windowed$-$pointwise J_p $+0.087$, win-rate 0.79, Wilcoxon $p = 10^{-33}$; energy $+0.052$; Table 2, RANDOM control 0.086). On real data, detection rises then dilutes in m (energy-net J_p, σ=0.05; no per-anomaly footprints identified): 0.409 (m=1), 0.494 (m=`slidingWindow`), 0.362 (m=400) (Table 15).

On the synthetic benchmark (480 configurations, known L), for short anomalies the measured optimum exceeds the raw anomaly duration: embedding length-w subsequences gives a raw anomaly

Table 3: Equal-treatment audit (TSB-AD-U): each baseline's own score window-averaged exactly as ours. Baseline rows: saved seed-2024 scores; ours: 3-seed means at σ=0.35. All metrics rise; LSTMAD on its 274 non-failing series. Gains are computed before rounding.

		VUS-PR			buffer-free AUC-PR		
Method	n	point	agg.	gain	point	agg.	gain
J_p (ours)	350	0.473	0.561	+0.087	0.372	0.490	+0.118
energy $= -\log p$ (ours)	349	0.518	0.570	+0.052	0.438	0.519	+0.080
J_{s_θ}, direct net (ours)	350	0.506	0.588	+0.081	0.416	0.521	+0.105
CNN	350	0.391	0.528	+0.137	0.376	0.469	+0.094
DTE	322	0.461	0.524	+0.063	0.365	0.447	+0.082
LSTMAD	274	0.287	0.520	+0.233	0.307	0.505	+0.198
TranAD	350	0.291	0.394	+0.103	0.293	0.360	+0.066
USAD	350	0.395	0.412	+0.017	0.362	0.381	+0.019
RANDOM	350	0.086	–	–	0.046	–	–

of length L a score footprint of roughly $\mathbf{L} + \mathbf{w} - \mathbf{1}$ positions. The supported statement is *scale*: on a dense 17-point m grid (buffer-free AUC-PR, Figure 1) the footprint is within 0.026 AUC-PR of the optimum for $L \geq 3$ (L=1: alignment-dominated, Appendix F), against pointwise rises of 0.045–0.414; the argmax equals it at L=3, 10 and undershoots on a near-flat curve at $L \geq 30$ (Tables 19–20). Estimator noise is also suppressed: the Hutchinson probe count has no measurable effect (windowed VUS-PR spread ≤ 0.0006 across $K = 1/4/16$/exact; Table 16).

An equal-treatment audit of the leaderboard. The identical moving average lifts all five baselines' saved scores by 0.017 to 0.233 (Table 3; USAD, internally smoothed, gains least; PaAno is patch-aggregated and not re-aggregated). The ordering changes, the field compresses (seed 2024); every lift persists on the common 246-series cohort (0.023 to 0.229; CNN and LSTMAD overtake USAD; Table 27) and under buffer-free AUC-PR (Table 3): the gains are not VUS buffering alone.

The limits: the inversion class on real flat-lines. Corollary 1 predicts a blind spot: anomalies sharper than normal data ($I(q) - I(p) > F(q\|p)$) invert the statistic; Figure 5 shows an inversion consistent with this class on a real low-variance SED flat-line (J below normal on both nets).

Per-source breakdown. Re-ranked against the twelve per-source columns, the direct net is 1st–3rd on six of the seven largest U sources; its one large deficit to PaAno is the long-anomaly OPPORTUNITY family (0.279 vs. 0.766; 9th) (Appendix D). We do not run the UCR Anomaly Archive separately: 233 of its 251 series are already in the TSB-AD-U release (content-matched).

Negative results. *(i) Two-sided and energy.* The two-sided $|T - \mathrm{med}(T_{\mathrm{cal}})|$ costs 0.014 VUS-PR (0.545 vs. 0.559, energy-net J_p, seed 2024), wins 20% of series, and forfeits conformal validity (Section 4.4); J_p does not surpass the same net's windowed energy (see above). *(ii) Score-norm / MSMA.* Adding divergence to the same network's score norm is worth +0.023 (component ablation, Table 1, $p = 0.025$); an architecture-matched MSMA-style surrogate (single-Gaussian auxiliary; Mahmood et al., 2021) reaches 0.550/0.514 (direct/energy net), 0.037 below the full statistic on the direct net (paired, seed 2024; bootstrap CI $[+0.008, +0.067]$, Wilcoxon $p = 0.063$; vs. the surrogate only, not original MSMA). *(iii) Conformal drift.* Naive split conformal over-alarms under drift (Barber et al., 2023): direct net, 3 seeds, exact rank, FAR mean 0.224 $\gg$ median 0.097 at α=0.01; rolling label-free recalibration cuts mean FAR to 0.053/0.099 at TPR cost 0.733→0.391 / 0.765→0.527, no formal guarantee (Table 29).

7 CONCLUSION

A score-only model shows no significant difference from the strongest baseline on TSB-AD-U (0.588 vs. 0.584 VUS-PR; PaAno leads under buffer-free AUC-PR) at under a third of its runtime; the divergence term adds a significant +0.023; and the gap $\frac{1}{2}F(q\|p) + \frac{1}{2}(I(p) - I(q))$ has a blind spot consistent with an observed flat-line inversion. Footprint-scale aggregation performs near the best tested scale and lifts every re-aggregated baseline.

REPRODUCIBILITY STATEMENT

Theory with regularity conditions is in Section 4 (proofs Appendix A); estimator, architecture, DSM objective, σ ladder, and Hutchinson settings in Section 3.2 and Appendix B.1. The protocol (VUS-PR; three-seed own-method summaries with the seed-2024 baseline aggregation audit and AxonAD convention; per-row coverage; the RANDOM baseline; the disjoint 48-series tuning split; the disclosed σ-selection researcher degree of freedom) is in Sections 5 and 6; full leaderboards, per-source breakdowns, the synthetic-benchmark specification, and the patched/failed-baseline log are in Appendices D–I.

THE USE OF LARGE LANGUAGE MODELS

In this work we used LLM-based tools in three capacities, consistent with the declarations in the submission form. (i) *Writing and drafting:* LLMs helped draft and condense parts of the manuscript and polish exposition; all text was reviewed and edited by the authors, and every quantitative statement was checked against the result files listed in the reproducibility statement. (ii) *Retrieval and discovery:* LLMs assisted the literature search and the related-work survey; retrieved references were read and verified by the authors before citation. (iii) *Research execution:* LLMs assisted in implementing experiment code (benchmark runners, ablations, analysis and figure scripts) and in stress-testing the experimental methodology and the mathematical statements; all code was audited, tested against closed-form cases, and its outputs verified against the released result files, and all mathematical claims and proofs were checked by the authors. We did not use generative AI to generate benchmark data (the synthetic benchmark uses conventional pseudo-random generators specified in Appendix F); translation and qualitative/thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and artifacts.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research (TMLR)*, 2024.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: UnSupervised anomaly detection on multivariate time series. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3395–3404, 2020.

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

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. SAND: Streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 14(10):1717–1729, 2021.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, 2009.

Wuxia Chen, Sean Moushegian, Vahid Tarokh, and Taposh Banerjee. Score-based quickest change detection and fault identification for multi-stream signals. *arXiv preprint arXiv:2511.03967*, 2025.

Yuhang Chen, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. ImDiffusion: Imputed diffusion models for multivariate time series anomaly detection. *Proceedings of the VLDB Endowment*, 17(3): 359–372, 2023.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, 2016.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024.

Utsav Dutta, Sina Khoshfetrat Pakazad, and Henrik Ohlsson. Time to embed: Unlocking foundation models for time series with channel descriptions. *arXiv preprint arXiv:2505.14543*, 2025.

Vijay Ekambaram, Subodh Kumar, Arindam Jati, Sumanta Mukherjee, Tomoya Sakai, Pankaj Dayama, Wesley M. Gifford, and Jayant Kalagnanam. TSPulse: Tiny pre-trained models with disentangled representations for rapid time-series analysis. In *International Conference on Learning Representations (ICLR)*, 2026.

Arik Ermshaus, Patrick Schäfer, and Ulf Leser. Window size selection in unsupervised time series analytics: A review and benchmark. In *Advanced Analytics and Learning on Temporal Data (AALTD 2022)*, volume 13812 of *Lecture Notes in Computer Science*. Springer, 2023. doi: 10.1007/978-3-031-24378-3_6.

Kamil Faber, Marcin Pietroń, Dominik Żurek, and Roberto Corizzo. xLSTMAD: A powerful xLSTM-based method for anomaly detection. In *IEEE International Conference on Data Mining (ICDM)*, 2025.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In *KI-2012: Poster and Demo Track*, pp. 59–63, 2012.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

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.

Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1479–1489, 2021.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9–10):1641–1650, 2003.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 387–395, 2018.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 19(2):433–450, 1990.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Dongmin Kim, Sunghyun Park, and Jaegul Choo. When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection. In *AAAI Conference on Artificial Intelligence*, 2024.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 7194–7201, 2022.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. In *International Conference on Machine Learning (ICML)*, 2026.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74 (1–2):67–94, 2015.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. V. N. Vishwanathan (eds.), *Predicting Structured Data*. MIT Press, 2006.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. COPOD: Copula-based outlier detection. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1118–1123, 2020.

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

Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, 2016.

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 (NeurIPS) Datasets and Benchmarks Track*, 2024.

Qinghua Liu and John Paparrizos. TSB-AD: benchmark code and bundled leaderboard results. https://github.com/TheDatumOrg/TSB-AD, 2026. Leaderboard CSVs in `benchmark_exp/leaderboard_results/`, commit `6beac72e`, accessed September 2026.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024.

Zongye Lyu. Did we actually fix it? An independent adversarial stress-test of post-point-adjustment evaluation metrics for time-series anomaly detection. *arXiv preprint arXiv:2607.11969*, 2026.

Ahsan Mahmood, Junier Oliva, and Martin Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations (ICLR)*, 2021.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 89–94, 2015.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2019.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019.

Kadir-Kaan Özer, René Ebeling, and Markus Enzweiler. Surprised by attention: Predictable query dynamics for time series anomaly detection. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Applied Data Science Track*, volume 16950 of *Lecture Notes in Computer Science*, pp. 217–234. Springer, 2026a.

Kadir-Kaan Özer, René Ebeling, and Markus Enzweiler. Just repair: A minimal denoising network for time series anomaly detection. *arXiv preprint arXiv:2604.17388*, 2026b.

Kadir-Kaan Özer, René Ebeling, and Markus Enzweiler. STREAM-VAE: Dual-path routing for slow and fast dynamics in vehicle telemetry anomaly detection. In *IEEE Intelligent Vehicles Symposium (IV)*, 2026c.

John Paparrizos and Luis Gravano. k-Shape: Efficient and accurate clustering of time series. In *ACM SIGMOD International Conference on Management of Data*, pp. 1855–1870, 2015.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022.

Jinju Park and Seokho Kang. PaAno: Patch-based representation learning for time-series anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2026.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD International Conference on Management of Data*, pp. 427–438, 2000.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at Microsoft. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3009–3017, 2019a.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019b.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Workshop on Machine Learning for Sensory Data Analysis (MLSDA)*, pp. 4–11, 2014.

M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulakis. Position: Quo vadis, unsupervised time series anomaly detection? In *International Conference on Machine Learning (ICML)*, 2024.

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.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations (ICLR)*, 2020.

Stephane Shao, Pierre E. Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the Hyvärinen score: Computation and consistency. *Journal of the American Statistical Association*, 114(528):1826–1837, 2019.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. In *IEEE Foundations and New Directions of Data Mining Workshop (in conjunction with ICDM)*, pp. 172–179, 2003.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2828–2837, 2019.

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.

George L. Turin. An introduction to matched filters. *IRE Transactions on Information Theory*, 6(3): 311–329, 1960.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023a.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023.

Suya Wu, Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh. Score-based quickest change detection for unnormalized models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10546–10565, 2023b.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations (ICLR)*, 2025.

Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 649–655, 2022.

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, et al. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *The World Wide Web Conference (WWW)*, pp. 187–196, 2018.

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.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations (ICLR)*, 2024.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3033–3045, 2023.

Chin-Chia Michael Yeh. Matrix profile for time-series anomaly detection: A reproducible open-source benchmark on TSB-AD. *arXiv preprint arXiv:2604.02445*, 2026.

Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1317–1322, 2016.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning (ICML)*, 2016.

Quan Zhou, Changhua Pei, Fei Sun, Jing Han, Zhengwei Gao, Dan Pei, Haiming Zhang, Gaogang Xie, and Jianhui Li. KAN-AD: Time series anomaly detection with Kolmogorov–Arnold networks. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79136–79149, 2025.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Full proofs

This appendix states the regularity assumptions and gives the full proofs of the separation identity, the Gaussian reduction and large-noise limit, the window concentration bound, and the split-conformal validity result.

Throughout, p denotes the normal (in-distribution) density on $\mathbb{R}^d$ and q an alternative (anomalous) density. We write $s_p := \nabla \log p$ and $s_q := \nabla \log q$ for the two score fields, and

$$J_p(x) \;:=\; \nabla^2 \log p(x) + \tfrac{1}{2}\|\nabla \log p(x)\|^2 \;=\; \nabla\cdot s_p(x) + \tfrac{1}{2}\|s_p(x)\|^2 \tag{9}$$

for the Hyvärinen score of p ($\nabla^2 = \nabla\cdot\nabla$, the Laplacian). The quantity is free of the normalizing constant of p: if $p = \tilde{p}/Z$ then $\nabla \log p = \nabla \log \tilde{p}$, so equation 9 depends only on the unnormalized model. This is the property the method exploits.

A.1 Regularity assumptions

We collect the assumptions under which the separation identity holds; they are the standard score-matching regularity conditions (Hyvärinen, 2005).

Assumption 1 (Regularity)**.**

(R1) $p, q > 0$ on $\mathbb{R}^d$ and are twice continuously differentiable (C^2), so that s_p, s_q, $\nabla\cdot s_p$ and $\nabla\cdot s_q$ exist pointwise.

(R2) (Finite moments) $\mathbb{E}_q\|s_q\|^2 < \infty$, $\mathbb{E}_q\|s_p\|^2 < \infty$, $\mathbb{E}_q\,|\nabla\cdot s_p| < \infty$, and $\mathbb{E}_q\,|\nabla\cdot s_q| < \infty$ (the last is the self-q moment used to evaluate $\mathbb{E}_q[J_q]$). For the detection-gap corollary we additionally assume the normal density p is regular in the same sense under itself: $I(p) = \mathbb{E}_p\|s_p\|^2 < \infty$ and $\mathbb{E}_p\,|\nabla\cdot s_p| < \infty$.

(R3) (Vanishing boundary flux) For every coordinate i, $q(x)\,s_{p,i}(x) \to 0$ and $q(x)\,s_{q,i}(x) \to 0$ as $\|x\| \to \infty$, and both $q\,s_p$ and $q\,s_q$ are integrable enough that the divergence theorem applies with no boundary term; equivalently $\int_{\mathbb{R}^d} \nabla\cdot\big(q\,s_p\big)\,dx = 0$ and $\int_{\mathbb{R}^d} \nabla\cdot\big(q\,s_q\big)\,dx = 0$. For the detection-gap corollary we likewise require the self-p flux condition $p(x)\,s_{p,i}(x) \to 0$ and $\int_{\mathbb{R}^d} \nabla\cdot\big(p\,s_p\big)\,dx = 0$.

The cross conditions ($q\,s_p$) justify Lemma 2 and the right-hand side $\mathbb{E}_q[J_p]$ of Theorem 1; the self-q conditions ($q\,s_q$) justify its left-hand side $\mathbb{E}_q[J_q]$; and the self-p conditions justify $\mathbb{E}_p[J_p] = -\frac{1}{2}I(p)$ used in Corollary 1.

Condition (R3) is where the anomalous density q must not place mass arbitrarily far out relative to the decay of the normal score s_p; it holds, for example, whenever q has sub-Gaussian tails and s_p grows at most polynomially, which covers the Gaussian and the learned energy models we use (the quadratic anchor of Appendix B.1 makes s_p grow linearly at infinity, so (R3) holds whenever q places negligible mass at the tails at which s_p is evaluated; z-scoring only recenters and rescales the windows and does not by itself guarantee light tails).

A.2 The integration-by-parts lemma

Lemma 2 (Hyvärinen's identity)**.** *Under Assumption 1,*

$$\mathbb{E}_q\big[s_q(x)^\top s_p(x)\big] \;=\; -\,\mathbb{E}_q\big[\nabla\cdot s_p(x)\big].$$

Proof. Write the left-hand side as an integral and use $q\,\nabla\log q = \nabla q$:

$$\mathbb{E}_q\big[s_q^\top s_p\big] = \int_{\mathbb{R}^d} q(x)\,\big(\nabla\log q(x)\big)^\top s_p(x)\,dx = \int_{\mathbb{R}^d} \big(\nabla q(x)\big)^\top s_p(x)\,dx.$$

Apply the divergence identity $\nabla\cdot\,(q\,s_p) = (\nabla q)^\top s_p + q\,(\nabla\cdot s_p)$ and integrate:

$$\int (\nabla q)^\top s_p\,dx = \int \nabla\cdot\,(q\,s_p)\,dx - \int q\,(\nabla\cdot s_p)\,dx.$$

By (R3) the first integral on the right vanishes (no boundary flux), leaving $\int(\nabla q)^\top s_p\,dx = -\int q\,(\nabla\cdot s_p)\,dx \;=\; -\mathbb{E}_q[\nabla\cdot s_p]$. Finiteness of every term is guaranteed by (R2), so the manipulation is justified. □

A.3 Proof of the exact separation identity (Theorem 1)

Theorem (Separation, $\frac{1}{2}$-Fisher; restatement of Theorem 1)**.** *Under Assumption 1,*

$$\boxed{\;\mathbb{E}_q\big[J_p\big] \;-\; \mathbb{E}_q\big[J_q\big] \;=\; \tfrac{1}{2}\,F(q\,\|\,p) \;\geq\; 0\;}$$

where $F(q\|p) := \mathbb{E}_{x\sim q}\big\|\nabla\log q(x) - \nabla\log p(x)\big\|^2$ is the Fisher divergence, with equality iff $q = p$ (q-a.e.).

Proof. Set $a := s_q$, $b := s_p$. Expand the Fisher divergence and apply Lemma 2 to the cross term:

$$\begin{aligned} F(q\|p) &= \mathbb{E}_q\|a\|^2 - 2\,\mathbb{E}_q[a^\top b] + \mathbb{E}_q\|b\|^2 \\ &= \mathbb{E}_q\|s_q\|^2 + 2\,\mathbb{E}_q[\nabla\cdot s_p] + \mathbb{E}_q\|s_p\|^2 \qquad \text{(Lemma 2)} \\ &= \mathbb{E}_q\|s_q\|^2 + 2\,\mathbb{E}_q\Big[\nabla\cdot s_p + \tfrac{1}{2}\|s_p\|^2\Big] = \mathbb{E}_q\|s_q\|^2 + 2\,\mathbb{E}_q[J_p]. \end{aligned}$$

Hence $\mathbb{E}_q[J_p] = \frac{1}{2}F(q\|p) - \frac{1}{2}\mathbb{E}_q\|s_q\|^2$. Applying the same integration-by-parts computation with q in place of p (using the self-q moment and flux conditions of Assumption 1, so $\mathbb{E}_q[s_q^\top s_q] = -\mathbb{E}_q[\nabla\cdot s_q]$ and $F(q\|q) = 0$) gives $\mathbb{E}_q[J_q] = -\frac{1}{2}\mathbb{E}_q\|s_q\|^2$. Subtracting eliminates the q-only term $-\frac{1}{2}\mathbb{E}_q\|s_q\|^2$:

$$\mathbb{E}_q[J_p] - \mathbb{E}_q[J_q] = \tfrac{1}{2}F(q\|p).$$

Nonnegativity is immediate since $F \geq 0$, and $F(q\|p) = 0$ iff $s_q = s_p$ q-a.e., i.e. $q = p$ up to a normalizing constant, which for two densities means $q = p$ (q-a.e.). □

The identity is *exact*, not asymptotic; the constant is $\frac{1}{2}$; and the "distance" is a genuine divergence. We verified it numerically to relative error $\leq 2.3 \times 10^{-3}$ on Gaussian q (`src/test_jp.py`, which checks the estimate against the closed form $\frac{1}{2}(\|\mu\|^2 + d(\sigma - 1/\sigma)^2)$ for $q = \mathcal{N}(\mu, \sigma^2 I)$, $p = \mathcal{N}(0, I)$).

What Theorem 1 does and does not license. The statement is about expectations under q, not about single points. It licenses *window-level* detection: if a window of m points is drawn from q, the window average $\frac{1}{m}\sum_i J_p(x_i)$ concentrates around $\mathbb{E}_q[J_q] + \frac{1}{2}F(q\|p)$, and *whenever the detection gap* $\Delta = \frac{1}{2}F(q\|p) + \frac{1}{2}(I(p) - I(q))$ *is nonzero* (and under the sub-exponential tail assumptions of Theorem 2) this value is separated from the normal-data value $\mathbb{E}_p[J_p]$ (Theorem 2). It does *not* certify that $J_p(x)$ is an optimal pointwise score; pointwise use is a heuristic the theory motivates but does not prove. This is why the window statistic is the headline and the pointwise statistic is secondary.

A.4 Gaussian reduction and the large-σ limit

Proposition 2 (Gaussian reduction). *If $p = \mathcal{N}(0, I_d)$, then $s_p(x) = -x$, $\nabla^2 \log p = -d$, and*

$$J_p(x) = \tfrac{1}{2}\|x\|^2 - d.$$

The statistic reduces to half the squared Euclidean (here also Mahalanobis) distance, up to the constant $-d$. More generally, for $p = \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, $s_p(x) = -\Sigma^{-1}(x - \mu)$, $\nabla\cdot s_p = -\operatorname{tr}(\Sigma^{-1})$, and $J_p(x) = \frac{1}{2}(x-\mu)^\top\Sigma^{-2}(x-\mu) - \operatorname{tr}(\Sigma^{-1})$.

Proof. For $p = \mathcal{N}(0, I_d)$, $\log p(x) = -\frac{1}{2}\|x\|^2 - \frac{d}{2}\log(2\pi)$, so $\nabla \log p = -x$ and $\nabla^2 \log p = \nabla\cdot(-x) = -d$. Substituting into equation 9 gives $J_p(x) = -d + \frac{1}{2}\|x\|^2$. For general $p = \mathcal{N}(\mu, \Sigma)$, $\log p(x) = -\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) + \text{const}$, so $s_p(x) = \nabla \log p = -\Sigma^{-1}(x-\mu)$ and $\nabla\cdot s_p = -\operatorname{tr}(\Sigma^{-1})$. Substituting into equation 9, $J_p(x) = -\operatorname{tr}(\Sigma^{-1}) + \frac{1}{2}\|\Sigma^{-1}(x-\mu)\|^2 = \frac{1}{2}(x-\mu)^\top\Sigma^{-2}(x-\mu) - \operatorname{tr}(\Sigma^{-1})$. □

Proposition 2 shows the statistic reduces to a quadratic form (Mahalanobis for $\Sigma = I$; Σ^{-2}-weighted in general) for Gaussian p, so any novelty must come from complex, learned, multimodal p, which is the regime our score networks target.

Corollary 3 (Large-noise limit, Gaussian family). *Let $p = \mathcal{N}(\mu, \Sigma)$ and let $p_\sigma = p * \mathcal{N}(0, \sigma^2 I) = \mathcal{N}(\mu, \Sigma + \sigma^2 I)$ be the DSM-smoothed model at noise scale σ. By Proposition 2, $J_{p_\sigma}(x) = \frac{1}{2}(x-\mu)^\top(\Sigma + \sigma^2 I)^{-2}(x-\mu) - \operatorname{tr}((\Sigma + \sigma^2 I)^{-1})$ exactly. Since $(\Sigma + \sigma^2 I)^{-2} = \sigma^{-4} I + O(\sigma^{-6})$, as $\sigma \to \infty$*

$$J_{p_\sigma}(x) = \tfrac{1}{2\sigma^4}\|x-\mu\|^2 - \tfrac{d}{\sigma^2} + O(\sigma^{-6}\|x-\mu\|^2) + O(\sigma^{-4}),$$

whose leading x-dependent term is a positive multiple (σ^{-4}) of the centered radial quadratic form $\|x-\mu\|^2$ (the larger $-d/\sigma^2$ offset is constant in x). The energy $-\log p_\sigma$ is affine in $(x-\mu)^\top(\Sigma + \sigma^2 I)^{-1}(x-\mu)$, whose leading x-dependent term is a positive multiple (σ^{-2}) of the same *form $\|x-\mu\|^2$: the two leading coefficients differ in their σ-power but not in the underlying quadratic form. Thus at large σ the Hyvärinen score and the energy $-\log p_\sigma$ are, after their respective positive σ-dependent rescalings, increasing affine functions of the same leading centered quadratic and hence agree in ranking away from ties.*

Proof. For Gaussian p, $p_\sigma = \mathcal{N}(\mu, \Sigma + \sigma^2 I)$ is exactly Gaussian, so Proposition 2 applies with covariance $\Sigma+\sigma^2 I$ and gives J_{p_σ} as the stated exact quadratic. Expanding $(\Sigma+\sigma^2 I)^{-1} = \sigma^{-2}(I+\Sigma/\sigma^2)^{-1} = \sigma^{-2}I - \sigma^{-4}\Sigma + O(\sigma^{-6})$ and squaring gives $(\Sigma+\sigma^2 I)^{-2} = \sigma^{-4}I + O(\sigma^{-6})$, so the leading quadratic is $\frac{1}{2\sigma^4}\|x-\mu\|^2$; $-\log p_\sigma$ is affine in $(x-\mu)^\top(\Sigma+\sigma^2 I)^{-1}(x-\mu)$ with leading term $\frac{1}{2\sigma^2}\|x-\mu\|^2$. The two leading coefficients carry different σ-powers (σ^{-4} vs. σ^{-2}) but the same centered quadratic form $\|x-\mu\|^2$, so after positive σ-dependent rescaling both induce the same ranking away from ties in the limit. For a general (non-Gaussian) p the same conclusion requires finite-moment/smoothness conditions and a uniform standardized expansion; we do not use that extension. □

This limit is not the operating regime. Empirically the ranking agreement between J_p and the energy does *not* increase toward 1 as σ grows over the tested range (median Spearman $0.898/0.905/0.892$ at $\sigma = 0.05/0.2/0.35$; Table 14); the two statistics remain distinct at every σ we use. We state Corollary 3 only as a limit and report explicitly that it is not reached.

A.5 Proof of the window-concentration bound (Theorem 2)

We first record the complexity of scoring, then the concentration bound.

Proposition 3 (Scoring complexity)**.** *With K Hutchinson probes, scoring B points costs $O\big(B(K+1)\big)$ score-network evaluations and $O(Bd)$ memory. The exact trace costs d backward passes per batch, i.e. $O(Bd)$ evaluations.*

Proof. Each point needs one forward/backward to obtain $s_p(x)$ (for $\|s_p\|^2$) and K Jacobian-vector products $v^\top \nabla s_p\, v$ for the Hutchinson estimate of $\nabla\cdot s_p = \mathrm{tr}(\nabla s_p)$, each a single extra backward pass; hence $K+1$ evaluations. The exact trace evaluates d coordinates of the Jacobian diagonal, i.e. d backward passes. Memory is dominated by storing the d-vector score for each of the B points. □

Theorem (Finite-sample window concentration; restatement and strengthening of Theorem 2)**.** *Let $x_1, \dots, x_m \overset{i.i.d.}{\sim} q$ and suppose $J_p(x)$ is sub-exponential under q with sub-exponential norm $\|J_p - \mathbb{E}_q J_p\|_{\psi_1} = b_q$ and variance proxy ν_q^2. Then there is a universal constant $c > 0$ such that for all $t > 0$,*

$$\Pr\Big[\Big|\tfrac{1}{m}\textstyle\sum_{i=1}^m J_p(x_i) - \mathbb{E}_q J_p\Big| > t\Big] \;\le\; 2\exp\Big(-c\,m\min\big(t^2/\nu_q^2,\; t/b_q\big)\Big).$$

Suppose in addition that J_p is sub-exponential under p with parameters (ν_p, b_p), and write $\nu := \max(\nu_p, \nu_q)$, $b := \max(b_p, b_q)$. Let

$$\Delta := \mathbb{E}_q[J_p] - \mathbb{E}_p[J_p] = \tfrac{1}{2}F(q\|p) + \tfrac{1}{2}\big(I(p) - I(q)\big)$$

be the detection gap of Corollary 1. Then, whenever $\Delta \neq 0$*, an anomalous window and a normal window each of length*

$$m \;\gtrsim\; \frac{1}{c}\,\max\Big\{\frac{\nu^2}{\Delta^2},\; \frac{b}{|\Delta|}\Big\}\,\log\frac{4}{\delta}$$

suffice to place both empirical means within $|\Delta|/4$ of their population values with confidence $1-\delta$, hence strictly *on opposite sides of the midpoint $\frac{1}{2}(\mathbb{E}_q[J_p] + \mathbb{E}_p[J_p])$ (each mean is then at least $|\Delta|/4 > 0$ from the midpoint).*

Proof. The first display is Bernstein's inequality for sums of i.i.d. sub-exponential random variables (Vershynin, 2018, Thm. 2.8.1), applied to $Y_i := J_p(x_i) - \mathbb{E}_q J_p$, which are centered and sub-exponential with the stated parameters. For the sample-complexity statement, the guarantee is conditional on $\Delta \neq 0$, and *no unconditional lower bound on Δ in terms of F alone exists* (for $p = \mathcal{N}(0,1)$, $q = \mathcal{N}(a, 1-a^2)$ with $0 < |a| < 1$ has $\Delta = \frac{1}{2}(a^2 + (1-a^2) - 1) = 0$ while $F > 0$), which is why the radius is set by $|\Delta|$ and not by F. Apply the first display under q (parameters ν_q, b_q) to the anomalous window mean and, by the added sub-exponentiality under p (parameters ν_p, b_p), the analogous Bernstein bound to the normal window mean, both with $t = |\Delta|/4$. Bounding each of the two failure probabilities by $\delta/2$ via the worst-case parameters $\nu = \max(\nu_p, \nu_q)$, $b = \max(b_p, b_q)$ and taking a union bound over the two events gives the stated m, where the two branches come from the sub-Gaussian (t^2/ν^2) and sub-exponential (t/b) regimes of the Bernstein

bound; both window means then lie within $|\Delta|/4$ of their population values. Since the two population values are exactly $|\Delta|$ apart and the midpoint is equidistant from each, each empirical mean is then within $|\Delta|/4$ of its own population value and hence at least $|\Delta|/2 - |\Delta|/4 = |\Delta|/4 > 0$ from the midpoint on the correct side; the two means are therefore strictly on opposite sides of the midpoint. □

This is a sample-complexity statement conditional on the detection gap: an explicit window length m sufficient to separate a gap of size Δ at confidence $1 - \delta$ for windows drawn purely from q and from p. The all-q/all-p model here contains no anomaly length, so the observed rise-then-dilute window-length curve (Table 15) is *not* predicted by this bound alone; that shape follows from the i.i.d. mixture SNR proxy stated next, under its additional assumptions.

The ρ-dilution argument (window longer than the anomaly). Theorem 2 assumes every point in the window is drawn from q. In practice a length-m window covering $\ell < m$ anomalous scores is a mixture: a fraction $\rho = \ell/m$ of anomalous points and $1 - \rho$ normal points, where $\ell \approx L+w-1$ is the score footprint of a raw anomaly of length L in the embedded pipeline (Section 4.2). The window mean then concentrates around the mixture mean

$$\rho\,\mathbb{E}_q[J_p] + (1-\rho)\,\mathbb{E}_p[J_p] \;=\; \mathbb{E}_p[J_p] + \rho\big(\mathbb{E}_q[J_p] - \mathbb{E}_p[J_p]\big),$$

so the *effective* separation shrinks by the dilution factor $\rho = \ell/m$. Increasing m beyond ℓ trades a smaller variance (the $1/m$ in Theorem 2) against a smaller signal (ρ falling as $1/m$), and the product $\rho^2 m \propto \ell^2/m$ *decreases*. Under independent fixed-p/q samples, a fixed nonzero gap Δ, and comparable tail scales, this SNR proxy rises for $m \le \ell$ and falls for $m > \ell$, so it peaks near $m^\star \approx \ell$: detection improves as the window matches the anomaly's footprint and falls off on either side. This is a proxy argument, not an exact optimizer in the correlated, embedded, heterogeneous time-series pipeline, and not a consequence of the all-q concentration bound (Theorem 2) on its own. Its predicted rise-then-dilute shape is confirmed both on real data (Table 15, peak at $m \approx$ `slidingWindow`) and on the synthetic benchmark with known L (Table 21: the argmax-m increases with L once the signal clears the noise floor, sitting above L for short anomalies because of the embedding footprint).

Remark 1 (Dependence). The i.i.d. assumption is idealized for a time series. If a dependent Bernstein-type inequality is available (for instance under β-mixing with a specified mixing rate and tail conditions, via a blocking argument that supplies an effective sample size $m_{\text{eff}} \approx m/\tau$ set by the mixing time τ, with the attendant constants and additive remainders), then the concentration bound and the Δ^{-2} scaling of the sample-complexity carry over with m_{eff} in place of m, and the mixture SNR proxy's qualitative rise-then-dilute behavior is unchanged under its own additional assumptions. We state this only as a conditional extension; β-mixing alone does not by itself yield the simple m/τ substitution.

Empirical dependence of the deployed statistic. To quantify how far the deployed pipeline is from the i.i.d. model of Theorem 2, we measure the dependence of the pointwise J series directly (headline direct net, seed 2024, U; series with $T \ge 1200$: 341 series at $m \in \{25, 100\}$, 300 at m=400; `dependence_U.csv`, `src/runtime_dependence.py`). The median lag-1 autocorrelation of pointwise J is 0.945, i.e. an AR(1) effective-sample-size ratio of ≈ 0.03. Comparing the observed standard deviation of disjoint length-m block means with the i.i.d. prediction $\mathrm{sd}(J)/\sqrt{m}$ gives median variance-inflation factors of 21 (m=25), 60 (m=100), and 107 (m=400): window means do concentrate with growing m, but far more slowly than the i.i.d. rate. Theorem 2's sample sizes should therefore be read with m replaced by an effective $m_{\text{eff}} \approx m/\mathrm{VIF}$ (roughly $m/20$ to $m/100$ on these series); the qualitative rise-then-dilute mechanism and the footprint scale are unaffected, and the conformal guarantee (which ranks window statistics rather than averaging them) depends on exchangeability, not on this inflation.

A.6 Proof of split-conformal validity (Theorem 3)

Theorem (Distribution-free false-alarm control; restatement of Theorem 3)**.** *Let T be the window statistic (e.g. $\frac{1}{m}\sum J_p$). Fix n calibration windows drawn from the normal distribution and, for a*

test window, define the conformal p-value

$$\pi \;=\; \frac{1+\big|\{\, i : T_i^{\mathrm{cal}} \geq T^{\mathrm{test}} \,\}\big|}{n+1}.$$

If the calibration windows and the test window are exchangeable*, then for any level* $\alpha \in (0,1)$*,* $\Pr[\pi \leq \alpha] \leq \alpha$*; declaring an alarm when* $\pi \leq \alpha$ *controls the false-alarm rate on normal windows at level* α*, regardless of the quality of the score model.*

Proof. This is the standard split-conformal argument (Vovk et al., 2005; Bates et al., 2023). Consider the augmented multiset $\{T_1^{\mathrm{cal}}, \ldots, T_n^{\mathrm{cal}}, T^{\mathrm{test}}\}$, which is exchangeable by hypothesis. Introduce i.i.d. $U(0,1)$ tie-breakers independent of the scores and define the randomized rank of T^{test} from the top with ties broken by these auxiliary variables; by exchangeability this randomized rank is uniform on $\{1, \ldots, n+1\}$, so the randomized p-value π_{rand} satisfies $\Pr[\pi_{\mathrm{rand}} \leq \alpha] \leq \alpha$. The deterministic p-value defined with "$\geq$", $\pi = \big(1+|\{i : T_i^{\mathrm{cal}} \geq T^{\mathrm{test}}\}|\big)/(n+1)$, counts every tie as an exceedance and therefore satisfies $\pi \geq \pi_{\mathrm{rand}}$ pointwise. Hence $\Pr[\pi \leq \alpha] \leq \Pr[\pi_{\mathrm{rand}} \leq \alpha] \leq \alpha$; that is, π is super-uniform (exactly uniform in the absence of ties). Declaring an alarm when $\pi \leq \alpha$ therefore controls the false-alarm rate at level α. The guarantee is on the calibration, not the model: it holds for any measurable T. □

Why it fails in practice on TSB-AD (a diagnosed negative result). Theorem 3 requires exchangeability between the calibration windows (from the training prefix) and the test windows (the remainder of the series). On real time series this fails: prefix and remainder are separated in time and subject to drift, so the two are not exchangeable. The observed pattern is consistent with the coverage-gap bound for conformal prediction beyond exchangeability (Barber et al., 2023): because the gap is *additive* on the false-alarm side, heterogeneous per-series additive gaps produce a realized FAR centered above α, right-skewed across series (mean $\gg$ median), and proportionally worst at small α. Empirically the statistic over-alarms. On the headline direct net (350 series, seeds $\{2024, 2025, 2026\}$, final-20%-of-prefix calibration, threshold the exact $\lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil$-th order statistic; no series' calibration set is too small at either α; median $n_{\mathrm{cal}} = 558$): at $\alpha = 0.05$ the mean false-alarm rate is 0.260 (median 0.137), about $5\times$ nominal, and at $\alpha = 0.01$ it is mean 0.224 (median 0.097), about $22\times$ nominal (Table 29; `conformal_v2_U.csv`). This shape is consistent with heterogeneous per-series additive gaps under non-exchangeability. Label-free rolling recalibration on a trailing, mostly-normal window (same exact-rank rule) reduces the realized false-alarm rate toward nominal (to mean 0.099 at $\alpha = 0.05$, 0.053 at $\alpha = 0.01$) at a cost in TPR, without a formal guarantee under arbitrary drift (Table 29). An alternative configuration (energy net, 50% calibration split, interpolated quantile; Table 28, `conformal_fix_U.csv`) gives the same qualitative picture; the procedure above yields moderately larger naive FARs.

A.7 Derivation of the plug-in bridge bound (Proposition 1)

Writing $e := s_\theta - s_{p_\sigma}$ and $J_{s_\theta} := \nabla \cdot s_\theta + \frac{1}{2}\|s_\theta\|^2$, expanding gives the pointwise decomposition

$$J_{s_\theta}(x) - J_{p_\sigma}(x) = \nabla \cdot e(x) + s_{p_\sigma}(x)^\top e(x) + \tfrac{1}{2}\|e(x)\|^2.$$

Assume e is C^1 and, for $r \in \{q, p_\sigma\}$, $I(r) < \infty$, $e \in L^2(r)$, $\mathbb{E}_r|\nabla \cdot e| < \infty$, and $r\, e$ has zero boundary flux; then $s_r^\top e$ is integrable (Cauchy–Schwarz on $I(r) < \infty$, $e \in L^2(r)$) and integrating the divergence term by parts term-by-term yields $\mathbb{E}_r[J_{s_\theta} - J_{p_\sigma}] = \mathbb{E}_r[(s_{p_\sigma} - s_r)^\top e] + \frac{1}{2}\mathbb{E}_r\|e\|^2$. With $F(q\|p_\sigma) < \infty$ the cross term is integrable, and Cauchy–Schwarz plus $\mathbb{E}_{p_\sigma}[(s_{p_\sigma} - s_{p_\sigma})^\top e] = 0$ give the bound equation 6.

A.8 Proof of the end-to-end operational gap (Corollary 2)

Recall $e := s_\theta - s_{p_\sigma}$. Under the assumptions of Proposition 1 for $r \in \{q, p\}$, the term-by-term expectation identity

$$\mathbb{E}_r[J_{s_\theta} - J_{p_\sigma}] = \mathbb{E}_r[(s_{p_\sigma} - s_r)^\top e] + \tfrac{1}{2}\,\mathbb{E}_r\|e\|^2$$

holds (the term-by-term expectation identity derived in Appendix A.7 by integrating the divergence term by parts under the zero-boundary-flux and integrability hypotheses). Applying it at $r = q$ and

$r = p$ and subtracting,

$$\big(\mathbb{E}_q[J_{s_\theta}] - \mathbb{E}_p[J_{s_\theta}]\big) - \Delta_{\mathrm{op}} = \mathbb{E}_q[(s_{p_\sigma} - s_q)^\top e] - \mathbb{E}_p[(s_{p_\sigma} - s_p)^\top e] + \tfrac{1}{2}\mathbb{E}_q\|e\|^2 - \tfrac{1}{2}\mathbb{E}_p\|e\|^2,$$

where $\Delta_{\mathrm{op}} = \mathbb{E}_q[J_{p_\sigma}] - \mathbb{E}_p[J_{p_\sigma}]$ is the population operational gap equation 7. Bound each cross term by Cauchy–Schwarz, using $\mathbb{E}_r\|s_{p_\sigma} - s_r\|^2 = F(r \,\|\, p_\sigma)$, and the triangle inequality to obtain equation 8. All four right-hand terms vanish when $s_\theta = s_{p_\sigma}$ (so $e = 0$), and the concentration statement of Theorem 2 applies to the deployed statistic J_{s_θ} with the deployed gap $\Delta_\theta := \mathbb{E}_q[J_{s_\theta}] - \mathbb{E}_p[J_{s_\theta}]$ in place of Δ.

B Implementation details

This appendix specifies the detector implementation: the energy network architecture, its denoising-score-matching training, the scoring statistics it returns, and the fidelity checks for the JuRe and DTE baselines.

All numbers in this section trace to `src/jp_detector.py` (code fingerprint `sha256/8 = fbbf24e1`, `results/MANIFEST.md`). The detector is built on TSB-AD's subsequence-embedding convention so that the detector window equals the pre-registered `slidingWindow`.

B.1 EnergyNet: architecture, training, scoring

Subsequence embedding. A univariate/multivariate series $x \in \mathbb{R}^{T\times f}$ is z-scored using *training-prefix statistics only* and embedded into overlapping length-w subsequences flattened to dimension $d = w \cdot f$, where $w = \texttt{slidingWindow} = \texttt{find_length_rank}(\text{channel } 0)$. Following TSB-AD's own convention, `find_length_rank` is computed on the *full* series (label-free autocorrelation period estimate): the period estimate is transductive, identical for every method, and the only full-series quantity in the pipeline (z-scoring and training remain train-prefix-only). This is the same embedding used by TSB-AD's `Sub_*` detectors, enforcing the parity rule (detector window = VUS buffer window).

Direct score parameterization (headline). The headline model is a directly parameterized score network $s_\theta(x, \sigma) \in \mathbb{R}^d$: an MLP with 3 hidden layers of width 256 and SiLU activations, input $[x;\, \mathrm{RFF}(\log\sigma)]$ with the same frozen 32-dimensional random-Fourier embedding of $\log\sigma$ as below, and a linear $\mathbb{R}^{256} \to \mathbb{R}^d$ output head plus a fixed linear anchor $-0.05\,x$ (the gradient of the energy net's quadratic anchor, so the two parameterizations share the same *explicit linear score anchor*; the learned component is unconstrained). No scalar energy head exists. Training, noise ladder, optimizer, and evaluation are identical to the energy network's below.

Energy parameterization. The second model is an energy network $E_\theta(x, \sigma)$ with score $s_\theta = -\nabla_x E_\theta$, so the energy $-\log p_\sigma \approx E_\theta$ is available from the *same* network for the key J_p-vs-energy ablation. The MLP body is 3 hidden layers of width 256 with SiLU activations. The noise scale is injected through a fixed random-Fourier embedding of $\log\sigma$ of dimension 32 ($\texttt{freqs}= \mathcal{N}(0, 4^2)$, frozen). A quadratic anchor $\frac{1}{2}\|x\|^2 \cdot (0.05)$ is added to the scalar head to keep the energy coercive far from the data, so the induced density is well posed and the score grows linearly at infinity (which helps satisfy (R3) of Assumption 1 when q has light tails).

Denoising score matching. Training minimizes the σ^2-weighted DSM loss

$$\mathcal{L} = \tfrac{1}{2}\big\|\sigma\, s_\theta(x + \sigma\varepsilon, \sigma) + \varepsilon\big\|^2, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

which is the standard DSM target $\nabla \log q_\sigma(\tilde{x}|x) = -\varepsilon/\sigma$ in σ-weighted form (Vincent, 2011; Song & Ermon, 2019). Noise scales are drawn from a geometric ladder of 8 scales on $[0.05, 1.0]$; the smallest scale is used at test time by default, with the operating σ selected on the tuning split (Appendix E). Optimization uses AdamW ($\texttt{lr}= 10^{-3}$, weight decay 10^{-5}) with a OneCycle schedule, batch size 256, gradient-norm clipping at 5, and a fixed budget of at least 20 epochs *and* at least 1500 optimizer updates (the step floor protects short prefixes). The final 20% of the (contiguous) normal windows are held out as the conformal calibration tail (kept contiguous so the calibration windows are drawn from the same regime as adjacent test windows; exchangeability is an assumption, not a consequence, of this choice). The code has three safeguards that alter this recipe on degenerate

prefixes (random subsampling to 200k training windows before the split, which would make the tail non-contiguous; calibrating on the training windows when fewer than 8 remain; shrinking the embedding window when the prefix yields fewer than 64 windows); none is triggered on any of the 530 evaluation series (verified from the run metadata).

Divergence estimator. $\nabla \cdot s_\theta = \mathrm{tr}(\nabla s_\theta)$ is computed *exactly* (d backward passes) when $d \leq 64$, and by Hutchinson's estimator with $K = 4$ Rademacher probes when $d > 64$. After window aggregation the choice of K has no measurable effect on the metric (Table 16: windowed VUS-PR $= 0.4969/0.4974/0.4974/0.4973$ for $K = 1/4/16/\text{exact}$, energy net at σ=0.05), so $K = 4$ is used throughout; the pointwise rank correlation to the exact trace rises with K ($0.88/0.93/0.96$), but the difference is removed by window averaging.

Statistics returned. From one forward evaluation the detector emits, per time step (padded to series length exactly as TSB-AD `Sub_*`): J_p pointwise (`jp_point`); the window average of J_p over $m =$ `slidingWindow` (`jp_window`, the headline); the ablations $\frac{1}{2}\|s\|^2$ (`sq_only`) and $\nabla \cdot s$ (`div_only`); the energy E_θ (`neglogp`); and the conformal p-value of the window statistic. An opt-in MSMA-style feature (log squared score norm across the noise ladder with a Gaussian fit) is provided for the ablation of Appendix D.

Aggregation. The headline aggregation length is $m =$ `slidingWindow`; the centered moving mean is edge-shrunk to preserve series length. The window-length sweep of Table 15 varies m over $\{1, 5, 25, \texttt{slidingWindow}, 100, 400\}$ ("sw" in the tables).

Timing. Mean end-to-end fit-plus-score time per TSB-AD-U series on a single L40S GPU is 3.0 s for the direct score net and 9.5 s for the energy net (seed 2024; `runtime_U.csv`, Appendix I).

B.2 JuRe re-implementation fidelity

Our JuRe port reproduces the published architecture: parameter counts 17,665 and 19,978 match the paper's two configurations exactly, and the TSB-AD-M AUC-PR of our port (0.404) equals the paper's reported 0.404. We therefore treat the port as faithful and report it as a baseline under the identical protocol (3 seeds, VUS-PR headline).

B.3 DTE adapter

DTE (Livernoche et al., 2024) is run from the official code with a thin window adapter that (i) feeds the same `slidingWindow` subsequence embedding used by all detectors and (ii) maps DTE's per-window density estimate to a per-time-step score with the identical padding convention. No changes are made to the DTE model itself; the adapter only aligns I/O to the shared protocol. DTE scores 0.4629 ± 0.0039 (U) and 0.3447 ± 0.0040 (M), 3 seeds.

C Architecture and training ablation (tuning split)

Table 4 sweeps the network and training configuration on the disjoint TSB-AD-U *tuning* split (48 series; the table basis is the 48 per-seed result files at seed 2024, one file per tuning series). We report the window-averaged J_p and window-averaged energy ($-\log p$) VUS-PR means. Two findings stand out. First, the convolutional energy is *clearly worse* than the MLP for both statistics (J_p 0.37–0.40 vs. $\geq$ 0.46 for every MLP row); this contrasts with JuRe's report that convolutional backbones help their reconstruction objective, and indicates the gain there does not transfer to the score-matching energy. Second, performance is *flat across MLP sizes and depths*: width $64 \to 512$ and depth $1 \to 5$ move J_p by at most ≈ 0.05, and the default (h=256, depth 3) is within noise of the best. The quadratic anchor and the exact optimizer-step budget make only small differences. This flatness is why we fix a single default configuration and do not tune the architecture on the evaluation set.

Table 4: Architecture/training ablation on the TSB-AD-U tuning split (window-averaged VUS-PR, seed 2024, evaluation σ=0.35). The convolutional energy is clearly worse; MLP results are flat across width and depth. Default row in the first line. The *params* column is the trainable parameter count at the representative univariate input dimension d=100 (window = 100, 1 feature, σ-embed 32); it varies with d but the ordering is fixed. The final row is the directly parameterized score network (no energy exists), whose entry in the J_p column is its window-aggregated Hyvärinen statistic J_{s_θ}; its field need not be conservative, and the plug-in bridge (Proposition 1) makes no conservativity assumption. The energy column ($-\log p$ point, 0.4696 for the default row) is the *pointwise* VUS-PR of the raw energy readout on the 48 tuning series; the "windowed energy" $\approx$ 0.519 of Table 11 is the moving-window aggregate on the 46 common series, so the two differ in aggregation and cohort. Source: `arch_ablation_U_tuning.csv`.

Configuration	params	n	J_p window	$-\log p$ point
default (MLP h=256, d3, ladder $8\times[.05, 1]$, anchor .05)	166k	48	0.5138	0.4696
MLP width h=64	17k	48	0.4646	0.4588
MLP width h=512	594k	48	0.5072	0.4569
MLP depth 1	34k	48	0.5142	0.4603
MLP depth 5	297k	48	0.4970	0.4568
Conv energy h=64	54k	48	0.3656	0.3763
Conv energy h=128	190k	48	0.4004	0.3943
No quadratic anchor	166k	48	0.4918	0.4680
500 optimizer steps	166k	48	0.4928	0.4482
5000 optimizer steps	166k	48	0.4889	0.4560
Noise ladder 3 scales	166k	48	0.4783	0.4585
Noise ladder $8 \times [0.02, 2]$	166k	48	0.4818	0.4590
Direct score net (MLP h=256, d3; no energy)	191k	48	0.5044	–

Table 5: Paired per-series comparisons on TSB-AD-U (VUS-PR): mean per-series difference, 95% bootstrap CI (10,000 resamples, seed 0), Wilcoxon signed-rank p. Rows 1–4 use per-series means over seeds $\{2024, 2025, 2026\}$; the MSMA rows pair on seed 2024 (MSMA single-seed). The CI (bootstrap of the mean) and the Wilcoxon test (signed ranks) can disagree, as in the MSMA-direct row. Source: `paired_tests_U.csv`.

Comparison	n	diff	95% CI	p
direct J vs. PaAno	350	+0.004	$[-0.032, +0.041]$	0.643
energy vs. J_p (energy net, σ=0.35)	349	+0.009	$[-0.017, +0.034]$	0.051
direct J vs. energy (σ=0.35)	349	+0.019	$[-0.005, +0.044]$	0.190
direct J vs. its score-norm (divergence gain)	350	+0.023	$[+0.007, +0.039]$	0.025
direct J vs. MSMA (direct, seed 2024)	350	+0.037	$[+0.008, +0.067]$	0.063
direct J vs. MSMA (energy net, seed 2024)	350	+0.074	$[+0.044, +0.104]$	$< 10^{-5}$

D FULL LEADERBOARDS AND PER-SOURCE BREAKDOWNS

Tables 6 and 8 give the expanded TSB-AD-U (350-series eval) and TSB-AD-M (180-series eval) baseline leaderboards under the strict protocol: VUS-PR headline with AUC-PR, VUS-ROC, AUC-ROC secondary, mean $\pm$ std over the pre-registered seeds $\{2024, 2025, 2026\}$, a RANDOM row in every table, and per-method n_{series} and n_{seeds}. The directly parameterized score net's headline J_{s_θ} (0.588 U / 0.427 M) and its component ablations are tabulated separately in Table 18; these tables carry its score-norm and divergence components (score norm and divergence, windowed). Rows with $n_{\text{seeds}} = 1$ are single-seed baselines (Time-RCD, AxonAD, MMPAD and the classical methods; see Appendix I); POLY and KMeansAD are TSB-AD's bundled implementations (Liu & Paparrizos, 2024). Rows with reduced n_{series} (< 350/< 180) are methods that failed on some series in TSB-AD's own code (documented in Appendix I); their means are computed over the series that ran, so they are not directly comparable to full-coverage rows and should be read with the coverage column. Rows for our own statistics name the network (direct score net or energy-parameterized net), the statistic (J_{s_θ}, J_p, energy, divergence $\nabla \cdot s$, score norm $\frac{1}{2}\|s\|^2$, MSMA-style surrogate), whether it is pointwise or windowed, and σ_{eval}; σ=0.05 is the DSM base scale (the energy at σ=0.05 covers 349 series after one TSB-AD crash). Partial-coverage σ-sweep diagnostics computed on tuning-split series ($n \leq 47$, 1–2 seeds) are *not* comparable to the full-coverage rows; they are set apart in Table 7. The component ordering differs between the two networks: in the energy net the windowed score

norm is weak (0.439 U at σ=0.35) and below the divergence (0.482), whereas in the direct net the windowed score norm is strong (0.565) and above the divergence (0.456); in both, the full statistic is best. The MSMA-style multiscale surrogate reaches 0.550 (direct net) and 0.514 (energy net) on U (seed 2024, 350 series), below the full Hyvärinen statistic of the same network in both cases.

Table 6: TSB-AD-U full leaderboard (350-series eval; VUS-PR headline, mean ± std over seeds; sorted by VUS-PR). Partial-coverage σ-sweep diagnostics are listed separately in Table 7.

Method	VUS-PR	AUC-PR	VUS-ROC	AUC-ROC	n	seeds
direct: J_{s_θ} (win., σ=0.35) (headline)	0.5879 ± 0.0009	0.5212 ± 0.0013	0.8691 ± 0.0004	0.8468 ± 0.0004	350	3
PaAno	0.5837 ± 0.0002	0.5516 ± 0.0001	0.9271 ± 0.0004	0.9016 ± 0.0002	350	3
energy net: energy (win., σ=0.05)	0.5797 ± 0.0057	0.5245 ± 0.0051	0.8730 ± 0.0045	0.8523 ± 0.0047	349	3
energy net: energy (win., σ=0.2)	0.5770 ± 0.0038	0.5231 ± 0.0050	0.8727 ± 0.0012	0.8522 ± 0.0011	350	3
energy net: energy (win., σ=0.35)	0.5704 ± 0.0022	0.5185 ± 0.0017	0.8717 ± 0.0018	0.8510 ± 0.0013	349	3
direct: score norm (win., σ=0.35)	0.5654 ± 0.0009	0.4999 ± 0.0013	0.8418 ± 0.0040	0.8214 ± 0.0039	350	3
energy net: J_p (win., σ=0.35)	0.5605 ± 0.0050	0.4900 ± 0.0057	0.8523 ± 0.0028	0.8263 ± 0.0022	350	3
direct: MSMA-style (win., multiscale)	0.5503	0.5008	0.8776	0.8535	350	1
energy net: J_p (win., σ=0.2)	0.5385 ± 0.0043	0.4693 ± 0.0031	0.8511 ± 0.0050	0.8260 ± 0.0042	350	3
Sub-KNN	0.5294	0.4437	0.8543	0.8292	350	1
energy net: energy (point, σ=0.05)	0.5230 ± 0.0057	0.4440 ± 0.0023	0.8694 ± 0.0038	0.8321 ± 0.0045	350	3
energy net: energy (point, σ=0.2)	0.5212 ± 0.0037	0.4399 ± 0.0028	0.8676 ± 0.0018	0.8281 ± 0.0031	350	3
energy net: divergence (win., σ=0.05)	0.5179 ± 0.0102	0.4684 ± 0.0101	0.8590 ± 0.0060	0.8353 ± 0.0069	347	3
energy net: energy (point, σ=0.35)	0.5172 ± 0.0019	0.4374 ± 0.0004	0.8654 ± 0.0022	0.8243 ± 0.0032	350	3
energy net: MSMA-style (win., multiscale)	0.5138	0.4561	0.8575	0.8322	350	1
direct: J_{s_θ} (point, σ=0.35)	0.5065 ± 0.0008	0.4158 ± 0.0024	0.8627 ± 0.0008	0.8122 ± 0.0011	350	3
energy net: J_p (win., σ=0.05)	0.4939 ± 0.0065	0.4187 ± 0.0079	0.8493 ± 0.0044	0.8253 ± 0.0046	350	3
Time-RCD	0.4907	0.4724	0.8794	0.8532	350	1
energy net: divergence (win., σ=0.35)	0.4824 ± 0.0148	0.4276 ± 0.0151	0.8024 ± 0.0126	0.7712 ± 0.0154	349	3
energy net: J_p (point, σ=0.35)	0.4731 ± 0.0078	0.3716 ± 0.0064	0.8466 ± 0.0027	0.7954 ± 0.0022	350	3
energy net: divergence (win., σ=0.2)	0.4681 ± 0.0145	0.4226 ± 0.0119	0.7805 ± 0.0091	0.7511 ± 0.0086	350	3
DTE	0.4629 ± 0.0039	0.3719 ± 0.0029	0.8505 ± 0.0030	0.8075 ± 0.0034	350	3
direct: divergence (win., σ=0.35)	0.4555 ± 0.0105	0.4090 ± 0.0094	0.7807 ± 0.0030	0.7557 ± 0.0029	350	3
energy net: divergence (point, σ=0.05)	0.4529 ± 0.0032	0.3712 ± 0.0029	0.8490 ± 0.0066	0.7963 ± 0.0090	350	3
energy net: J_p (point, σ=0.2)	0.4518 ± 0.0052	0.3502 ± 0.0046	0.8468 ± 0.0036	0.7967 ± 0.0039	350	3
AxonAD	0.4476	0.3882	0.8311	0.7751	350	1
StreamVAE	0.4441 ± 0.0110	0.3979 ± 0.0085	0.8327 ± 0.0065	0.7917 ± 0.0074	318	3
energy net: score norm (win., σ=0.35)	0.4391 ± 0.0119	0.3785 ± 0.0093	0.7487 ± 0.0211	0.7217 ± 0.0229	349	3
JuRe	0.4347 ± 0.0110	0.3978 ± 0.0113	0.7814 ± 0.0108	0.7567 ± 0.0126	350	3
Sub-MCD	0.4307	0.3453	0.8300	0.7937	347	1
energy net: divergence (point, σ=0.35)	0.4197 ± 0.0051	0.3376 ± 0.0021	0.8113 ± 0.0064	0.7465 ± 0.0088	350	3
Sub-PCA	0.4171	0.3771	0.7585	0.7144	350	1
energy net: divergence (point, σ=0.2)	0.4164 ± 0.0103	0.3426 ± 0.0024	0.7976 ± 0.0098	0.7347 ± 0.0109	350	3
energy net: J_p (point, σ=0.05)	0.4087 ± 0.0050	0.3003 ± 0.0056	0.8444 ± 0.0036	0.7986 ± 0.0034	350	3
USAD	0.3959 ± 0.0017	0.3631 ± 0.0013	0.7492 ± 0.0016	0.7027 ± 0.0021	350	3
FITS	0.3902 ± 0.0070	0.3128 ± 0.0053	0.7628 ± 0.0020	0.6834 ± 0.0040	350	3
POLY	0.3867	0.3125	0.7582	0.7266	350	1
CNN	0.3815 ± 0.0082	0.3656 ± 0.0089	0.8157 ± 0.0034	0.7466 ± 0.0039	350	3
KShapeAD	0.3798	0.3416	0.7355	0.7156	339	1
TimesNet	0.3707 ± 0.0044	0.2899 ± 0.0019	0.7421 ± 0.0016	0.6773 ± 0.0023	350	3
energy net: score norm (point, σ=0.35)	0.3678 ± 0.0082	0.2770 ± 0.0074	0.7497 ± 0.0165	0.6948 ± 0.0201	350	3
energy net: score norm (win., σ=0.2)	0.3612 ± 0.0028	0.3004 ± 0.0041	0.6716 ± 0.0086	0.6393 ± 0.0101	350	3
KMeansAD	0.3577	0.3180	0.7492	0.7373	350	1
Sub-HBOS	0.3489	0.3129	0.7469	0.7084	350	1
Sub-IForest	0.3457	0.3110	0.7839	0.7241	350	1
MatrixProfile	0.3449	0.2583	0.7622	0.7337	350	1
AutoEncoder	0.3444 ± 0.0031	0.2735 ± 0.0035	0.7935 ± 0.0007	0.7636 ± 0.0004	350	3
IForest	0.3434	0.3135	0.7657	0.7079	350	1
Sub-LOF	0.3314	0.2445	0.7677	0.7205	350	1
OmniAnomaly	0.3265 ± 0.0003	0.3233 ± 0.0007	0.7567 ± 0.0008	0.6858 ± 0.0005	350	3
SR	0.3208	0.3157	0.8082	0.7417	350	1
MMPAD	0.3098	0.2573	0.6371	0.6479	10	1
energy net: score norm (point, σ=0.2)	0.3031 ± 0.0039	0.2181 ± 0.0029	0.6811 ± 0.0121	0.6233 ± 0.0140	350	3
LSTMAD	0.2923 ± 0.0048	0.3118 ± 0.0044	0.7812 ± 0.0013	0.7213 ± 0.0024	274	3
TranAD	0.2870 ± 0.0036	0.2889 ± 0.0038	0.7183 ± 0.0037	0.6458 ± 0.0032	350	3
energy net: score norm (win., σ=0.05)	0.2733 ± 0.0092	0.2146 ± 0.0055	0.5982 ± 0.0078	0.5629 ± 0.0084	348	3
Sub-OCSVM	0.2357	0.1672	0.6276	0.5681	350	1
energy net: score norm (point, σ=0.05)	0.2327 ± 0.0055	0.1555 ± 0.0028	0.6122 ± 0.0074	0.5559 ± 0.0078	350	3
SAND	0.2217	0.2104	0.7050	0.6931	191	1
AnomalyTransformer	0.1028 ± 0.0026	0.0643 ± 0.0040	0.5517 ± 0.0026	0.4834 ± 0.0022	350	3
RANDOM	0.0858 ± 0.0003	0.0455 ± 0.0002	0.6271 ± 0.0027	0.5086 ± 0.0054	350	3

Table 7: Partial-coverage σ-sweep diagnostics, TSB-AD-U (*non-comparable* to the full-coverage leaderboard in Table 6). These rows are windowed-energy, divergence, and score-norm variants of the energy net evaluated only on tuning-split series ($n \leq 47$, 1–2 seeds); they exist to show the σ trend, not to rank against full-coverage methods. Sorted by VUS-PR.

Method	VUS-PR	AUC-PR	VUS-ROC	AUC-ROC	n	seeds
energy net: energy (win., σ=0.35)	0.5499 ± 0.0523	0.4522 ± 0.0414	0.8733 ± 0.0713	0.8632 ± 0.0797	3	2
energy net: energy (win., σ=0.2)	0.5387 ± 0.0303	0.4850 ± 0.0077	0.8356 ± 0.0408	0.8225 ± 0.0404	36	2
energy net: energy (win., σ=0.1)	0.5331 ± 0.0520	0.4841 ± 0.0208	0.8292 ± 0.0439	0.8176 ± 0.0462	41	2
energy net: divergence (win., σ=0.35)	0.5300 ± 0.0938	0.4066 ± 0.0254	0.8489 ± 0.1157	0.8292 ± 0.1383	3	2
energy net: energy (win., σ=0.05)	0.5249 ± 0.0086	0.4826 ± 0.0071	0.8264 ± 0.0038	0.8065 ± 0.0032	45	2
energy net: energy (win., multiscale)	0.5192	0.4793	0.8384	0.8183	46	1
energy net: energy (win., σ=0.27)	0.5188	0.4799	0.8313	0.8098	46	1
energy net: energy (win., σ=0.45)	0.5029	0.4808	0.8394	0.8212	46	1
energy net: energy (win., σ=0.6)	0.4922	0.4578	0.8229	0.8045	47	1
energy net: energy (win., σ=1.0)	0.4906	0.4496	0.8277	0.8082	47	1
energy net: divergence (win., σ=0.05)	0.4888 ± 0.0049	0.4350 ± 0.0126	0.8446 ± 0.0216	0.8164 ± 0.0221	45	2
energy net: divergence (win., multiscale)	0.4578	0.3978	0.7714	0.7371	46	1
energy net: divergence (win., σ=0.1)	0.4566 ± 0.0208	0.4017 ± 0.0101	0.7872 ± 0.0240	0.7644 ± 0.0286	41	2
energy net: divergence (win., σ=0.6)	0.4462	0.3732	0.7411	0.7103	47	1
energy net: divergence (win., σ=0.45)	0.4442	0.3621	0.7638	0.7351	46	1
energy net: divergence (win., σ=0.27)	0.4388	0.3571	0.7499	0.7136	46	1
energy net: score norm (win., σ=1.0)	0.4366	0.4059	0.7474	0.7224	47	1
energy net: score norm (win., σ=0.45)	0.4358	0.4053	0.7438	0.7209	46	1
energy net: divergence (win., σ=0.2)	0.4298 ± 0.0151	0.3829 ± 0.0092	0.7562 ± 0.0066	0.7300 ± 0.0060	36	2
energy net: score norm (win., σ=0.6)	0.4259	0.4033	0.7549	0.7385	47	1
energy net: divergence (win., σ=1.0)	0.3973	0.3218	0.7356	0.7080	47	1
energy net: score norm (win., σ=0.2)	0.3786 ± 0.0296	0.3234 ± 0.0108	0.6983 ± 0.0103	0.6749 ± 0.0211	36	2
energy net: score norm (win., σ=0.35)	0.3436 ± 0.0263	0.3313 ± 0.0144	0.7943 ± 0.2141	0.7800 ± 0.2248	3	2
energy net: score norm (win., σ=0.27)	0.3363	0.3208	0.6800	0.6599	46	1
energy net: score norm (win., multiscale)	0.3143	0.2911	0.6198	0.5909	46	1
energy net: score norm (win., σ=0.1)	0.3040 ± 0.0962	0.2743 ± 0.0498	0.6350 ± 0.0283	0.6165 ± 0.0200	41	2
energy net: score norm (win., σ=0.05)	0.2450 ± 0.0586	0.1988 ± 0.0537	0.5986 ± 0.0142	0.5674 ± 0.0177	45	2

Table 8: TSB-AD-M full leaderboard (180-series eval; VUS-PR headline, mean ± std over seeds; sorted by VUS-PR). AxonAD-M is the published #1 method ($\approx$ 0.49); its complete-attempt run under seed 2024 scores 0.494 on the 162/180 series that complete in its own code (matching the published 0.495; 18 in-code failures, Appendix I); an interrupted seed-2025 re-run (120 series, 0.487) is not pooled. On that identical 162-series subset our windowed energy (σ=0.2) scores 0.505 versus AxonAD's 0.494, the energy-net J_p scores 0.480, and the direct-net J_{s_θ} scores 0.472 (`axonad_M.csv`).

Method	VUS-PR	AUC-PR	VUS-ROC	AUC-ROC	n	seeds
AxonAD	0.4941	0.4416	0.8545	0.8240	162	1
energy net: energy (win., σ=0.05)	0.4676 ± 0.0024	0.4404 ± 0.0020	0.7817 ± 0.0066	0.7681 ± 0.0074	180	3
energy net: energy (win., σ=0.2)	0.4575 ± 0.0005	0.4334 ± 0.0016	0.7750 ± 0.0033	0.7610 ± 0.0034	180	3
energy net: energy (win., σ=0.35)	0.4541 ± 0.0029	0.4295 ± 0.0023	0.7745 ± 0.0016	0.7606 ± 0.0025	180	3
energy net: energy (point, σ=0.05)	0.4522 ± 0.0014	0.4051 ± 0.0005	0.7826 ± 0.0069	0.7582 ± 0.0070	180	3
energy net: energy (point, σ=0.2)	0.4426 ± 0.0013	0.3974 ± 0.0008	0.7744 ± 0.0033	0.7499 ± 0.0034	180	3
energy net: energy (point, σ=0.35)	0.4375 ± 0.0032	0.3936 ± 0.0012	0.7727 ± 0.0027	0.7484 ± 0.0028	180	3
energy net: J_p (win., σ=0.35)	0.4361 ± 0.0051	0.4127 ± 0.0049	0.7780 ± 0.0067	0.7701 ± 0.0076	180	3
direct: score norm (win., σ=0.35)	0.4354 ± 0.0032	0.4045 ± 0.0032	0.7608 ± 0.0004	0.7449 ± 0.0007	180	3
energy net: J_p (win., σ=0.05)	0.4328 ± 0.0052	0.4050 ± 0.0069	0.7657 ± 0.0028	0.7548 ± 0.0034	180	3
energy net: J_p (win., σ=0.2)	0.4305 ± 0.0085	0.4040 ± 0.0096	0.7676 ± 0.0176	0.7589 ± 0.0182	180	3
StreamVAE	0.4282 ± 0.0008	0.3866 ± 0.0023	0.7715 ± 0.0017	0.7554 ± 0.0018	180	3
direct: J_{s_θ} (win., σ=0.35) (headline)	0.4272 ± 0.0016	0.3968 ± 0.0017	0.7715 ± 0.0019	0.7600 ± 0.0021	180	3
energy net: J_p (point, σ=0.35)	0.4254 ± 0.0040	0.3766 ± 0.0054	0.7722 ± 0.0085	0.7424 ± 0.0086	180	3
energy net: J_p (point, σ=0.05)	0.4230 ± 0.0044	0.3704 ± 0.0059	0.7613 ± 0.0028	0.7331 ± 0.0029	180	3
JuRe	0.4227 ± 0.0026	0.4039 ± 0.0037	0.7591 ± 0.0035	0.7515 ± 0.0034	180	3
direct: J_{s_θ} (point, σ=0.35)	0.4213 ± 0.0013	0.3671 ± 0.0017	0.7640 ± 0.0008	0.7335 ± 0.0008	180	3
energy net: J_p (point, σ=0.2)	0.4189 ± 0.0087	0.3696 ± 0.0086	0.7661 ± 0.0148	0.7371 ± 0.0160	180	3
direct: divergence (win., σ=0.35)	0.4063 ± 0.0119	0.3694 ± 0.0126	0.7636 ± 0.0018	0.7424 ± 0.0032	180	3
PaAno	0.4050 ± 0.0001	0.3991 ± 0.0001	0.8062 ± 0.0000	0.7836 ± 0.0000	180	3
MMPAD	0.4033	0.3425	0.7836	0.7480	139	1
USAD	0.4028 ± 0.0018	0.3617 ± 0.0008	0.7658 ± 0.0011	0.7364 ± 0.0006	180	3
energy net: divergence (win., σ=0.05)	0.4012 ± 0.0070	0.3639 ± 0.0060	0.7572 ± 0.0079	0.7402 ± 0.0086	180	3
TimesNet	0.3874 ± 0.0022	0.3358 ± 0.0016	0.7090 ± 0.0023	0.6681 ± 0.0016	180	3
energy net: divergence (point, σ=0.05)	0.3861 ± 0.0017	0.3325 ± 0.0015	0.7580 ± 0.0055	0.7246 ± 0.0058	180	3
energy net: divergence (win., σ=0.2)	0.3816 ± 0.0077	0.3444 ± 0.0076	0.7236 ± 0.0165	0.7053 ± 0.0185	180	3
energy net: score norm (win., σ=0.35)	0.3782 ± 0.0167	0.3482 ± 0.0192	0.6937 ± 0.0123	0.6787 ± 0.0130	180	3
energy net: divergence (point, σ=0.2)	0.3658 ± 0.0052	0.3130 ± 0.0049	0.7310 ± 0.0137	0.6948 ± 0.0148	180	3
energy net: divergence (win., σ=0.35)	0.3639 ± 0.0111	0.3263 ± 0.0115	0.7192 ± 0.0119	0.7004 ± 0.0131	180	3
FITS	0.3623 ± 0.0002	0.3033 ± 0.0006	0.6991 ± 0.0004	0.6503 ± 0.0005	180	3
energy net: score norm (point, σ=0.35)	0.3585 ± 0.0184	0.3070 ± 0.0192	0.6977 ± 0.0134	0.6676 ± 0.0127	180	3

Continued on next page

Table 8 – continued

Method	VUS-PR	AUC-PR	VUS-ROC	AUC-ROC	n	seeds
CNN	0.3532 ± 0.0003	0.3513 ± 0.0010	0.8071 ± 0.0010	0.7738 ± 0.0016	180	3
DTE	0.3447 ± 0.0040	0.2883 ± 0.0037	0.7719 ± 0.0064	0.7387 ± 0.0054	180	3
energy net: divergence (point, σ=0.35)	0.3406 ± 0.0073	0.2910 ± 0.0072	0.7221 ± 0.0118	0.6822 ± 0.0125	180	3
energy net: score norm (win., σ=0.2)	0.3245 ± 0.0069	0.2959 ± 0.0087	0.6475 ± 0.0018	0.6329 ± 0.0022	180	3
TranAD	0.3184 ± 0.0007	0.3197 ± 0.0002	0.7463 ± 0.0008	0.7118 ± 0.0004	180	3
OmniAnomaly	0.3153 ± 0.0001	0.3152 ± 0.0001	0.7438 ± 0.0006	0.7064 ± 0.0007	180	3
energy net: score norm (point, σ=0.2)	0.3047 ± 0.0071	0.2546 ± 0.0075	0.6556 ± 0.0037	0.6260 ± 0.0044	180	3
AutoEncoder	0.2979 ± 0.0029	0.2978 ± 0.0017	0.6932 ± 0.0020	0.6751 ± 0.0019	180	3
LSTMAD	0.2978 ± 0.0015	0.3029 ± 0.0015	0.7628 ± 0.0052	0.7304 ± 0.0060	147	3
KMeansAD	0.2858	0.2458	0.7209	0.6829	180	1
energy net: score norm (win., σ=0.05)	0.2828 ± 0.0058	0.2529 ± 0.0062	0.5763 ± 0.0051	0.5561 ± 0.0035	180	3
PCA	0.2691	0.2409	0.7058	0.6759	180	1
energy net: score norm (point, σ=0.05)	0.2662 ± 0.0057	0.2184 ± 0.0048	0.5867 ± 0.0006	0.5529 ± 0.0014	180	3
OCSVM	0.2628	0.2354	0.6696	0.6341	180	1
CBLOF	0.2603	0.2648	0.6979	0.6686	179	1
IForest	0.2487	0.2095	0.7437	0.7040	180	1
RobustPCA	0.2187	0.2179	0.6017	0.5730	157	1
EIF	0.2074	0.1865	0.7075	0.6697	180	1
COPOD	0.2015	0.2046	0.6820	0.6518	180	1
HBOS	0.1856	0.1610	0.6691	0.6331	180	1
KNN	0.1807	0.1404	0.5729	0.4954	164	1
LOF	0.1392	0.1005	0.5922	0.5354	161	1
AnomalyTransformer	0.1175 ± 0.0031	0.0759 ± 0.0041	0.5088 ± 0.0052	0.4696 ± 0.0043	180	3
RANDOM	0.0957 ± 0.0002	0.0505 ± 0.0001	0.5785 ± 0.0020	0.4998 ± 0.0024	180	3

Per-source breakdowns. Tables 9 and 10 report VUS-PR by data source (mean over seeds; the energy-parameterized network's J_p at σ=0.35) alongside the same network's windowed energy (σ=0.35) and the strongest baselines, with the rank of J_p among all listed methods. The aggregate picture is not driven by one family: on U, J_p ranks 1–3 on 4 of the 7 largest sources (SMD 1st, WSD 1st, YAHOO 2nd, UCR 3rd; NAB 4th, Exathlon 7th, OPPORTUNITY 9th). For the energy-net J_p the gap to PaAno concentrates on the long-anomaly families OPPORTUNITY (J_p 0.204 vs. PaAno 0.766), SVDB and MITDB (ECG), matching the characterization of Appendix F. The headline *direct* score network (aggregate 0.588) shows the same per-source pattern but with a stronger top end. Recomputing the "ours" column with the direct net's window-aggregated J_{s_θ} (mean over the three seeds, `deepdive_long_U.csv`) and re-ranking it in place of the energy-net J_p among the same twelve method columns gives, on the seven largest sources, UCR 0.450 (rank 1), SMD 0.833 (rank 1), YAHOO 0.698 (rank 2), Exathlon 0.942 (rank 3), OPPORTUNITY 0.279 (rank 9), NAB 0.525 (rank 2), and WSD 0.620 (rank 2); the remaining sizeable sources are SVDB 0.635 (rank 2), SMAP 0.813 (rank 3), IOPS 0.402 (rank 4), and TODS 0.639 (rank 4). The direct net is thus 1–3 on six of the seven largest sources and, as with the energy-net J_p, carries its one large deficit on the long-anomaly OPPORTUNITY family (unlike the energy-net J_p, it leads PaAno on SVDB); the columns for the other methods are unchanged, so the energy-net J_p column above is retained for direct comparison.

Table 9: Per-source VUS-PR, TSB-AD-U (energy-net J_p, σ=0.35; mean over seeds; OPPORT. = OPPORTUNITY; T-RCD = Time-RCD, StrVAE = StreamVAE, S- = Sub-). The final column is the rank of J_p among all displayed method columns, including the RANDOM column; on near-noise sources such as `Power` a random scorer can outrank J_p, which this convention reflects.

source	n	J_p	energy	PaAno	T-RCD	AxonAD	StrVAE	JuRe	DTE	S-KNN	S-PCA	USAD	Random	rank
UCR	70	0.405	0.396	0.416	0.326	0.152	0.138	0.244	0.344	0.420	0.124	0.104	0.009	3
SMD	33	0.845	0.818	0.532	0.438	0.707	0.634	0.761	0.745	0.790	0.442	0.544	0.025	1
YAHOO	30	0.760	0.446	0.684	0.877	0.177	0.122	0.337	0.290	0.262	0.155	0.103	0.039	2
Exathlon	30	0.915	0.937	0.927	0.920	0.958	0.967	0.836	0.862	0.879	0.892	0.936	0.110	7
OPPORT.	27	0.204	0.541	0.766	0.575	0.669	0.734	0.491	0.158	0.101	0.914	0.798	0.070	9
NAB	23	0.505	0.512	0.437	0.369	0.532	0.509	0.407	0.423	0.487	0.395	0.495	0.125	4
WSD	20	0.656	0.650	0.507	0.242	0.282	0.122	0.228	0.360	0.516	0.123	0.032	0.014	1
SVDB	18	0.393	0.585	0.493	0.196	0.458	0.508	0.299	0.504	0.661	0.522	0.442	0.047	9
SMAP	17	0.743	0.698	0.854	0.441	0.603	0.504	0.543	0.685	0.834	0.501	0.463	0.037	3
IOPS	15	0.441	0.478	0.427	0.434	0.266	0.233	0.292	0.288	0.379	0.248	0.153	0.042	2
TODS	13	0.607	0.611	0.779	0.879	0.546	0.573	0.606	0.643	0.637	0.539	0.549	0.461	6
Stock	8	0.723	0.763	0.676	0.607	0.597	0.613	0.611	0.628	0.666	0.763	0.620	0.585	3
NEK	8	0.766	0.802	0.755	0.783	0.589	0.696	0.435	0.538	0.777	0.887	0.724	0.113	5
MGAB	8	0.014	0.009	0.198	0.044	0.041	0.359	0.004	0.098	0.281	0.005	0.006	0.004	7
LTDB	8	0.482	0.504	0.629	0.357	0.622	0.638	0.450	0.467	0.596	0.556	0.550	0.247	8
MSL	7	0.653	0.582	0.479	0.412	0.487	0.523	0.585	0.556	0.688	0.507	0.483	0.100	2
MITDB	7	0.202	0.329	0.313	0.314	0.229	0.207	0.211	0.214	0.502	0.310	0.217	0.059	11
SED	2	0.142	0.064	0.930	0.408	0.039	0.482	0.060	0.783	0.734	0.031	0.035	0.077	6
TAO	2	0.924	0.929	0.948	0.924	0.916	0.914	0.926	0.928	0.921	0.930	0.928	0.906	7

Continued on next page

source	n	J_p	energy	PaAno	T-RCD	AxonAD	StrVAE	JuRe	DTE	S-KNN	S-PCA	USAD	Random	rank
CATSv2	1	0.298	0.277	0.370	0.133	0.428	0.464	0.340	0.384	0.358	0.265	0.319	0.053	8
SWaT	1	0.311	0.472	0.181	0.219	0.529	0.316	0.313	0.500	0.098	0.393	0.355	0.121	8
Daphnet	1	0.541	0.533	0.423	0.233	0.379	0.493	0.468	0.547	0.521	0.422	0.446	0.063	2
Power	1	0.070	0.076	0.150	0.101	0.069	0.102	0.074	0.067	0.063	0.076	0.165	0.090	9

Table 10: Per-source VUS-PR, TSB-AD-M (energy-net J_p, σ=0.35; mean over seeds; AxonAD: seed 2024, series completing in its own code; OPPORT. = OPPORTUNITY; StrVAE = StreamVAE).

source	n	J_p	energy	PaAno	AxonAD	StrVAE	JuRe	DTE	KNN	PCA	USAD	Random	rank
SVDB	28	0.479	0.535	0.451	0.492	0.472	0.415	0.515	0.058	0.103	0.449	0.063	4
Exathlon	25	0.913	0.913	0.885	0.972	0.917	0.917	0.550	0.330	0.520	0.966	0.098	5
SMAP	25	0.451	0.508	0.356	0.354	0.491	0.508	0.427	0.075	0.295	0.266	0.035	4
GHL	23	0.027	0.015	0.011	0.465	0.023	0.019	0.020	0.007	0.037	0.013	0.011	3
SMD	20	0.261	0.383	0.343	0.415	0.361	0.297	0.225	0.282	0.340	0.396	0.042	9
MSL	14	0.445	0.434	0.296	0.336	0.440	0.466	0.180	0.162	0.375	0.306	0.077	2
MITDB	11	0.267	0.209	0.261	0.198	0.187	0.186	0.199	0.053	0.051	0.133	0.034	1
TAO	11	0.794	0.784	0.865	0.736	0.709	0.707	0.754	0.712	0.642	0.737	0.708	2
OPPORT.	7	0.185	0.164	0.149	0.267	0.133	0.130	0.144	0.056	0.214	0.394	0.047	4
CATSv2	5	0.301	0.254	0.087	0.279	0.116	0.251	0.155	0.071	0.061	0.186	0.033	1
LTDB	4	0.410	0.403	0.506	0.403	0.434	0.415	0.486	0.192	0.423	0.490	0.192	7
SWaT	2	0.333	0.346	0.310	0.425	0.383	0.388	0.259	0.115	0.365	0.452	0.140	7
CreditCard	1	0.138	0.114	0.042	0.094	0.062	0.073	0.089	0.024	0.030	0.064	0.020	1
PSM	1	0.194	0.197	0.178	–	0.196	0.248	0.202	0.121	0.170	0.169	0.126	5
Genesis	1	0.678	0.612	0.863	0.729	0.249	0.817	0.084	0.037	0.116	0.049	0.011	4
GECCO	1	0.692	0.166	0.085	0.185	0.044	0.051	0.375	0.110	0.036	0.123	0.022	1
Daphnet	1	0.419	0.439	0.299	0.324	0.331	0.332	0.224	0.246	0.112	0.126	0.063	2

E NOISE-SCALE SWEEP AND RANK AGREEMENT

This appendix reports the noise-scale selection on the tuning split, the fair J_p-versus-energy comparison on the evaluation sets, the per-series rank agreement between the two statistics, and the window-length and Hutchinson-K sweeps.

Tuning-split σ selection. Table 11 gives the complete noise-scale sweep on the disjoint TSB-AD-U tuning split (48 series, 46 common pairs). The grid is $\{0.05, 0.1, 0.2, 0.27, 0.35, 0.45, 0.6, 1.0\}$ plus a multiscale (σ-averaged) configuration. The raw windowed-J_p criterion peaks at σ=0.35 (0.5122, 0.0015 above the runner-up σ=0.45); the "fair" criterion (gap to the same net's windowed energy) peaks at σ=0.6 (+0.0125, 0.0195 above the -0.0070 gap at σ=0.35). We take σ=0.35 as the headline. The multiscale row does not beat the best single σ (0.489 vs. 0.512), consistent with the negative result that averaging over scales removes the σ hyperparameter at a cost of ≈ 0.02.

Table 11: TSB-AD-U tuning-split σ sweep (windowed VUS-PR, seed 2024; 46 series common to all configurations; `tuning_sigma.csv`). The multiscale row averages J_p over $\sigma \in \{0.1, 0.2, 0.35, 0.6\}$.

σ_{eval}	windowed J_p	windowed $-\log p$	gap	win-rate	n
multiscale	0.4890	0.5192	−0.0302	0.2174	46
σ=0.05	0.4452	0.5187	−0.0735	0.2174	46
σ=0.1	0.4623	0.5191	−0.0567	0.2391	46
σ=0.2	0.4960	0.5203	−0.0243	0.2826	46
σ=0.27	0.4888	0.5188	−0.0300	0.4130	46
σ=0.35	0.5122	0.5192	−0.0070	0.3913	46
σ=0.45	0.5107	0.5029	0.0079	0.4565	46
σ=0.6	0.5105	0.4980	0.0125	0.4565	46
σ=1.0	0.4842	0.4965	−0.0123	0.3913	46

Fair evaluation-set comparison. Tables 12 and 13 give the fair J_p-vs-energy comparison on the evaluation sets, both statistics window-averaged, all 3 seeds. On U the gap closes from -0.085 (σ=0.05) to -0.009 (σ=0.35); on M from -0.035 to -0.018. At the headline σ=0.35, J_p is within 0.01 (U) / 0.02 (M) of the same network's windowed energy using only the score, which is the case that matters for diffusion and score models with no accessible energy. The multiscale row appears in the tuning table above; the `div` and $\|s\|^2$ columns are the two additive components of J_p.

Table 12: Fair J_p-vs-energy comparison, TSB-AD-U (both window-averaged, 3 seeds); means over series–seed pairs, so they can differ in the fourth decimal from the per-seed means of Table 6). "gap" is $J_p - (-\log p)$; "win" is per-series win-rate of J_p; "$-\log p$ (pt)" is the *pointwise* (unaggregated) energy readout, shown for reference; all other columns are window-averaged. "pairs" counts series–seed pairs and can fall below 350×3 when individual series–seed runs fail; n counts series present in at least one seed. Significance is assessed per series in the main text (headline σ=0.35: per-series Wilcoxon $p = 0.051$, $n = 349$).

σ	pairs	n	seeds	J_p	$-\log p$ (win)	$-\log p$ (pt)	gap	win	div	$\|s\|^2$
0.05	1047	350	3	0.4950	0.5797	0.5240	−0.0846	0.3276	0.5179	0.2733
0.20	1050	350	3	0.5385	0.5770	0.5212	−0.0385	0.3886	0.4681	0.3612
0.35	1047	349	3	0.5618	0.5704	0.5183	−0.0086	0.4355	0.4824	0.4391

Table 13: Fair J_p-vs-energy comparison, TSB-AD-M (both window-averaged, 3 seeds); means over series–seed pairs). Significance is assessed per series in the main text.

σ	pairs	n	seeds	J_p	$-\log p$ (win)	$-\log p$ (pt)	gap	win	div	$\|s\|^2$
0.05	540	180	3	0.4328	0.4676	0.4522	−0.0348	0.3278	0.4012	0.2828
0.20	540	180	3	0.4305	0.4575	0.4426	−0.0270	0.3685	0.3816	0.3245
0.35	540	180	3	0.4361	0.4541	0.4375	−0.0180	0.3926	0.3639	0.3782

Rank agreement. Table 14 shows that the per-series Spearman rank correlation between J_p and the energy stays near 0.90 at *every* σ (median Spearman $0.898/0.905/0.892$), and this does not increase with σ. The two statistics therefore remain distinct across the operating range; the large-σ ranking limit of Corollary 3 is not reached.

Table 14: Per-series Spearman rank agreement between windowed J_p and windowed energy, TSB-AD-U.

suite	network	σ	Spearman mean	Spearman median	n
U	energy net	0.05	0.829	0.898	350
U	energy net	0.20	0.746	0.905	350
U	energy net	0.35	0.722	0.892	350

Window-length sweep and Hutchinson K. For completeness we repeat the two backbone ablations here with full precision. Table 15 is the window-length sweep on U (3 seeds): detection rises with m to a peak at $m \approx$ `slidingWindow` (0.4939) and falls for larger m (0.3621 at m=400), the rise-then-dilute shape of the mixture SNR proxy (Appendix A.5). Table 16 shows the Hutchinson probe count K has no measurable effect on the windowed metric.

Table 15: Window-length sweep on TSB-AD-U (energy-network windowed J_p, σ=0.05; VUS-PR mean ± std over 3 seeds, 350 series). Among the tested m, performance peaks at the per-series `slidingWindow` and decreases under excessive smoothing, consistent with the mixture SNR argument (Appendix A.5); the sweep does not identify per-anomaly footprints or establish an optimal aggregation length.

m	mean	std	seeds	n
m=1	0.4087	0.0050	3	350
m=5	0.4289	0.0043	3	350
m=25	0.4645	0.0056	3	350
m=sw	0.4939	0.0065	3	350
m=100	0.4693	0.0106	3	350
m=400	0.3621	0.0060	3	350

Table 16: Hutchinson K sweep (U, seed 2024, energy net, σ=0.05). Windowed VUS-PR was insensitive to the tested probe counts.

K	Sp. vs. exact	win. VUS-PR	n
K=1	0.8798	0.4969	350
K=4 (main)	0.9277	0.4974	350
K=16	0.9616	0.4974	350
exact	1.0000	0.4973	350

E.1 Further analyses: equal treatment, two-sided statistic, direct score net

This subsection collects three further analyses: an equal-treatment aggregation control, the two-sided window statistic, and the directly parameterized score network.

Equal-treatment aggregation. To show that window aggregation is not a privilege we grant only to our own statistic, we apply the *same* moving-average window to each baseline's own per-point score and re-evaluate. The full point $\rightarrow$ aggregated comparison under both the headline VUS-PR and the buffer-free AUC-PR control is the merged main-text Table 3 (seed 2024): every baseline improves by $+0.017$ to $+0.233$ VUS-PR, but their ordering changes (CNN becomes first among these five, LSTMAD overtakes USAD and TranAD) and none overtakes our windowed statistic. PaAno, which is already internally aggregated, is excluded. This is the equal-footing version of the "aggregate before you alarm" claim.

Two-sided vs. one-sided window statistic. The upper-tail guarantee holds when $I(q) \leq I(p)$ (and is inverted only when $I(q) - I(p) > F(q\|p)$); a two-sided statistic $|T - \mathrm{med}(T_{\mathrm{cal}})|$ is an exploratory variant aimed at negative-gap ($\Delta < 0$) anomalies that an upper-tail test misses, at a measured cost. Its center is the median of T over the training prefix (written $\mathrm{med}(T_{\mathrm{cal}})$ because the prefix contains the calibration windows it later ranks against); this breaks the exchangeability that Theorem 3 requires; we therefore report it as a sensitivity analysis without the split-conformal validity guarantee, not as a theory-safe default. Table 17 summarizes the comparison on U (350 series): the one-sided statistic scores 0.559 (energy-net J_p, σ=0.35, seed 2024; its 3-seed mean is 0.561), the two-sided variant 0.545, and the two-sided variant wins on only 19.7% of series. We therefore keep one-sided as the benchmark config and report two-sided as an exploratory option with its cost stated.

Table 17: One-sided vs. two-sided window statistic, TSB-AD-U (350 series, energy-net J_p, σ=0.35, seed 2024). Source: `twosided_U.csv`.

Statistic	one-sided	two-sided	two-sided win-rate
windowed VUS-PR (mean over 350 series)	0.559	0.545	0.197

Directly parameterized score network. The setting that motivates the score-only statistic is a model with no accessible energy. We therefore train a directly parameterized score network $s_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ (the architecture class of diffusion models), for which $-\log p$ is undefined, and score it with the window-aggregated Hyvärinen statistic J_{s_θ}; its field need not be conservative, and the plug-in bridge (Proposition 1) makes no conservativity assumption. Table 18 gives its tuning-split σ sweep (peaking at σ=0.35, matching the energy net's carried-over choice) and its evaluation-set result: 0.588 VUS-PR on U (350 series, 3 seeds) at σ=0.35, at the level of the energy-parameterized network's windowed energy (0.577 at σ=0.2) and PaAno (0.584), and above the energy-parameterized J_p (0.561 at σ=0.35); on M it reaches 0.427 ± 0.002 (180 series, 3 seeds).

Table 18: Direct score net (no energy exists). Top block: tuning-split windowed J_{s_θ} VUS-PR by σ (seed 2024). Middle: evaluation-set windowed J_{s_θ} VUS-PR (3 seeds). Bottom: component ablations of the same network on TSB-AD-U (sources: `direct_tuning_sigma.csv`, `leaderboard_U.csv`, `ablations_summary_{U,M}.csv`).

Split / σ	VUS-PR	n
tuning, σ=0.35	0.504	48
tuning, σ=0.2	0.498	48
tuning, σ=0.05	0.487	48
tuning, σ=0.6	0.477	48
U eval, σ=0.35 (3 seeds)	0.588 ± 0.001	350
M eval, σ=0.35 (3 seeds)	0.427 ± 0.002	180
U, score-norm only (windowed, 3 seeds)	0.565 ± 0.001	350
U, divergence only (windowed, 3 seeds)	0.456 ± 0.011	350
U, pointwise J (no aggregation, 3 seeds)	0.506 ± 0.001	350

F SYNTHETIC BENCHMARK WITH KNOWN GROUND TRUTH

This appendix describes the controlled synthetic benchmark and uses it to confirm the window-length prediction and the complementarity between the score-based statistic and the energy, with a matching characterization on real data.

Interaction between embedding length w and aggregation length m. The pipeline embeds length-w subsequences before scoring, so a raw anomaly of length L produces anomalous *scores* at roughly $L + w - 1$ positions. If dilution operates on this score footprint rather than on L, the optimal aggregation length should follow w when $L \ll w$ and L when $L \gg w$. Re-running the synthetic benchmark at $w \in \{25, 50, 100\}$ (2 seeds for $w \neq 50$) shows the short-anomaly trend, with finite-grid/finite-sample exceptions at larger L (buffer-free AUC-PR, J_p):

Table 19: Optimal aggregation length $m^\star$ by anomaly length L and embedding length w (grid $m \in \{1, 3, 10, 30, 100, 300\}$). For $L \leq 30$ the optimum increases with the embedding length (w=25/50 $\rightarrow$ 30, w=100 $\rightarrow$ 100), consistent with a score footprint of $\approx L + w - 1$ positions. At L=100 every w gives 100; at L=300 the optimum is 300 for $w \in \{50, 100\}$ but 100 for w=25. The table supports the short-anomaly footprint trend, not an exact $m^\star = L + w - 1$ law (`synthetic_wm.csv`).

	$w = 25$	$w = 50$	$w = 100$
$L = 1$	30	30	100
$L = 3$	30	30	100
$L = 10$	30	30	100
$L = 30$	30	30	100
$L = 100$	100	100	100
$L = 300$	100	300	300

Dense aggregation sweep. To test the footprint prediction beyond the coarse six-point grid, we re-ran the full synthetic benchmark (4 processes $\times$ 5 anomaly types $\times$ 6 lengths $\times$ 4 seeds $=$ 480 configurations, w=50, σ=0.35) on a 17-point aggregation grid $m \in \{1, 3, 10, 20, 30, 40, 50, 52, 59, 79, 100, 130, 149, 200, 300, 349, 400\}$ that includes every footprint value L+w−1; raw pointwise scores are stored so any future re-aggregation is free (`src/synthetic_dense.py`, `synthetic_dense_long.csv`, 8,160 rows). Table 20 reports, per length: the grid argmax of mean buffer-free AUC-PR, the per-seed argmax spread, the footprint value, and the margin by which the argmax beats the footprint. The supported conclusion is scale, not an exact law: the footprint is within 0.026 AUC-PR of the optimum for $L \geq 3$, *is* the argmax at L=3, 10, and is undershot by the argmax on a near-flat curve at $L \geq 30$. The L=1 argmax should not be read as a scale preference: a post-hoc fine sweep ($m \in \{49, \ldots, 53\}$, same stored scores) oscillates 0.120–0.200 between neighboring m, an alignment artifact of the centered moving mean (asymmetric by one sample for even m) interacting with one-point anomalies; the 0.053 margin at L=1 is within this oscillation.

Table 20: Dense sweep (buffer-free AUC-PR, mean over 80 configurations per length): grid argmax vs. footprint L+w−1 (w=50). "per-seed argmax" lists the argmax for each seed; "margin" is the AUC-PR by which the argmax beats the footprint. Source: `synthetic_dense_long.csv`.

L	argmax m	per-seed argmax	footprint	AUC-PR at argmax	margin
1	52	52, 52, 52, 52	50	0.173	+0.053
3	52	50, 50, 52, 52	52	0.304	0.000
10	59	59, 52, 59, 59	59	0.603	0.000
30	59	79, 59, 40, 59	79	0.716	+0.026
100	100	100, 79, 100, 100	149	0.778	+0.007
300	149	200, 149, 149, 149	349	0.793	+0.006

Specification. The controlled benchmark (`src/synthetic_bench.py`, fingerprint `c839cc51`) uses 4 normal processes (`ar2`, `seasonal`, `seasonal_ar`, `mixture`; the last is regime-switching between two seasonal patterns, i.e. a multimodal normal), 5 anomaly types (`spike`, `level_shift`, `variance`, `pattern_loss`, `freq_shift`), and a length grid $L \in \{1, 3, 10, 30, 100, 300\}$. Each test series has 3 planted anomalies, total length 12,000 with a 6,000-point clean training prefix and 4 seeds; this is $4 \times 5 \times 6 \times 4 = 480$ configurations. For each series one energy net is fit (σ_{eval}=0.35, window $=$ period or 50) and evaluated for $m \in \{1, 3, 10, 30, 100, 300\}$ with VUS-PR using `slidingWindow`$= L$ (the true anomaly length) as buffer.

Optimal window vs. true length (mixture SNR proxy, Appendix A.5). Table 21 reports windowed-J_p VUS-PR as a function of true length L (rows) and window m (columns), with the argmax-m in the last column. Aggregation trades variance reduction against dilution, and empirically the optimal aggregation length increases with anomaly extent: $m^\star = 30$ for $L \leq 30$ and $m^\star = 100$ for $L \geq 100$. The optimum sits *above* L for short anomalies rather than at L exactly, because the pipeline embeds length-w subsequences before scoring, so one raw anomaly of length L has a score footprint of roughly $L+w-1$ positions; with the synthetic $w \approx 50$ this predicts $m^\star$ well above L for $L \leq 10$, which is what the argmax column shows ($m^\star = 30$ at $L \in \{1, 3, 10\}$). The w-by-m interaction study (Table 19) shows this short-anomaly footprint trend directly by varying w.

Table 21: Synthetic optimal-window study (windowed J_p VUS-PR, averaged over processes, types, and seeds). Rows: true length L; columns: window m; last column: $\arg\max_m$.

True length L	m=1	m=3	m=10	m=30	m=100	m=300	$\arg\max m$
1	0.0148	0.0124	0.0116	0.0169	0.0080	0.0027	30
3	0.0605	0.0617	0.0619	0.0910	0.0459	0.0148	30
10	0.2214	0.2286	0.2568	0.3843	0.1853	0.0355	30
30	0.6307	0.6417	0.6615	0.7168	0.6753	0.1704	30
100	0.7644	0.7670	0.7724	0.7848	0.7931	0.6087	100
300	0.7929	0.7943	0.7959	0.7994	0.8015	0.8005	100

Buffer-free AUC-PR reanalysis. The optimal-window result above uses VUS-PR with a length-L buffer. To rule out any buffer artifact we repeat the study with buffer-free AUC-PR (no VUS smoothing, no buffer) for both statistics; Tables 22 and 23 give the result. *These tables are the original coarse six-point grid* ($m \in \{1, 3, 10, 30, 100, 300\}$); the dense 17-point sweep of Table 20 (identical generators, planting, embedding w=50, and seeds; only the m grid differs) supersedes them for locating optima, and the two agree exactly at shared grid values (e.g. L=30, m=30: 0.6589 in both). On the coarse grid the J_p argmax is nondecreasing in L and coincides with L at $30/100/300$; the dense grid shows those coincidences were grid artifacts (dense argmax $59/100/149$). For the windowed energy the coarse trend is the same up to a near-tie at L=30 (m=10: 0.7541 vs. m=30: 0.7521). At $L \in \{1, 3\}$ absolute AUC-PR is ≤ 0.11 and the argmax is noise-dominated; at L=10 the signal clears the floor (AUC-PR 0.379 at m=30 on this grid).

Table 22: Buffer-free AUC-PR, windowed J_p on the synthetic benchmark (averaged over processes, types, seeds). Rows: true length L; columns: window m; last column $\arg\max_m$. Coarse six-point grid; superseded for optima by the dense sweep of Table 20. Only $L \in \{1, 3\}$ are near the noise floor.

True length L	m=1	m=3	m=10	m=30	m=100	m=300	$\arg\max m$
1	0.0384	0.0246	0.0212	0.0416	0.0194	0.0218	30
3	0.0768	0.0838	0.0796	0.1067	0.0468	0.0117	30
10	0.1892	0.1959	0.2174	0.3790	0.1595	0.0254	30
30	0.5634	0.5776	0.5990	0.6589	0.6286	0.1195	30
100	0.6878	0.6940	0.7037	0.7299	0.7779	0.4498	100
300	0.7472	0.7500	0.7549	0.7659	0.7889	0.7894	300

Table 23: Buffer-free AUC-PR, windowed energy ($-\log p$) on the synthetic benchmark. Same layout as Table 22. The L=30 argmax falls at m=10 (0.7541) rather than 30 (0.7521), a near-tie on this coarse grid; the trend is otherwise nondecreasing.

True length L	m=1	m=3	m=10	m=30	m=100	m=300	$\arg\max m$
1	0.0224	0.0216	0.0266	0.0372	0.0153	0.0106	30
3	0.0698	0.0713	0.0840	0.0880	0.0509	0.0132	30
10	0.2048	0.2090	0.2215	0.3892	0.1594	0.0285	30
30	0.7495	0.7517	0.7541	0.7521	0.6002	0.1465	10
100	0.8291	0.8301	0.8321	0.8387	0.8506	0.4923	100
300	0.8337	0.8344	0.8357	0.8394	0.8505	0.8507	300

Complementarity by type and length. Table 24 gives the best-m VUS-PR per statistic for each (type, L) cell together with the J_p-minus-energy gap. The pattern: short anomalies often favor

J_p ($L \leq 10$, gap $+0.001$ to $+0.07$ on most types; exceptions `spike` L=10, gap -0.055, and `variance` L=1, gap -0.005), while the energy wins for long anomalies; the worst case for J_p is a `freq_shift` of length 300 (gap -0.222), a long *in-support* pattern change, which is exactly the regime the mechanism argument predicts the zeroth-order energy should win. This is the ground-truth counterpart of the real-data OPPORTUNITY/SVDB weakness.

Table 24: Synthetic type$\times L$ complementarity (best-m VUS-PR per statistic). Last column is J_p minus energy ($-\log p$); positive favors J_p.

Type	L	div	J_p	$-\log p$	$\|s\|^2$	$J_p - (-\log p)$
freq_shift	1	0.008	0.008	0.001	0.015	0.007
freq_shift	3	0.005	0.012	0.009	0.013	0.003
freq_shift	10	0.126	0.202	0.191	0.183	0.011
freq_shift	30	0.387	0.432	0.507	0.359	−0.075
freq_shift	100	0.514	0.513	0.693	0.504	−0.180
freq_shift	300	0.530	0.530	0.753	0.530	−0.222
level_shift	1	0.023	0.034	0.026	0.037	0.007
level_shift	3	0.120	0.155	0.116	0.131	0.039
level_shift	10	0.527	0.564	0.540	0.476	0.024
level_shift	30	0.886	0.894	0.975	0.703	−0.080
level_shift	100	0.934	0.973	0.998	0.773	−0.025
level_shift	300	0.941	0.990	1.000	0.825	−0.010
pattern_loss	1	0.008	0.008	0.001	0.015	0.007
pattern_loss	3	0.002	0.008	0.003	0.009	0.005
pattern_loss	10	0.051	0.094	0.093	0.090	0.001
pattern_loss	30	0.421	0.449	0.457	0.399	−0.008
pattern_loss	100	0.512	0.500	0.566	0.477	−0.066
pattern_loss	300	0.572	0.514	0.563	0.500	−0.049
spike	1	0.041	0.037	0.035	0.036	0.002
spike	3	0.152	0.160	0.139	0.147	0.021
spike	10	0.660	0.559	0.615	0.433	−0.055
spike	30	0.976	0.934	0.988	0.764	−0.053
spike	100	0.996	0.993	0.999	0.852	−0.006
spike	300	0.999	0.996	0.999	0.900	−0.003
variance	1	0.012	0.012	0.017	0.012	−0.005
variance	3	0.094	0.123	0.097	0.113	0.025
variance	10	0.492	0.502	0.436	0.360	0.066
variance	30	0.948	0.921	0.929	0.734	−0.007
variance	100	0.999	0.998	0.998	0.843	0.000
variance	300	1.000	1.000	1.000	0.870	0.000

Real-data characterization. Tables 25 and 26 give the same complementarity on real TSB-AD data at the headline σ=0.35 (both statistics windowed, 3 seeds), bucketed by anomaly type, by anomaly-length-over-window, by average anomaly length, and by source. On U, J_p beats the energy by $+0.377$ on point-only series and $+0.209$ for average length ≤ 5, and loses by -0.106 for length > 400 and -0.337 on OPPORTUNITY, which is the point and short versus long and regime-shift split that the synthetic study reproduces with ground truth.

Table 25: Real-data characterization, TSB-AD-U (J_p minus energy, both windowed, σ=0.35, 3 seeds; positive favors J_p).

split	bucket	n	mean gap	win-rate
type	mixed	21	−0.028	0.381
type	point-only	28	0.377	0.714

Continued on next page

split	bucket	n	mean gap	win-rate
type	sequence-only	300	−0.043	0.417
length/window	< 0.5	94	0.124	0.574
length/window	0.5–1	54	0.018	0.463
length/window	1–2	52	−0.046	0.481
length/window	2–5	52	−0.072	0.385
length/window	> 5	97	−0.097	0.299
avg. length	≤ 5	50	0.209	0.580
avg. length	6–25	49	0.008	0.531
avg. length	26–100	66	0.025	0.545
avg. length	101–400	102	−0.067	0.431
avg. length	> 400	82	−0.106	0.220
source	CATSv2	1	0.021	1.000
source	Daphnet	1	0.008	1.000
source	Exathlon	30	−0.022	0.233
source	IOPS	15	−0.036	0.600
source	LTDB	8	−0.023	0.250
source	MGAB	8	0.005	0.750
source	MITDB	6	−0.110	0.500
source	MSL	7	0.071	0.571
source	NAB	23	−0.007	0.304
source	NEK	8	−0.037	0.375
source	OPPORT.	27	−0.337	0.037
source	Power	1	−0.005	0.000
source	SED	2	0.078	1.000
source	SMAP	17	0.044	0.412
source	SMD	33	0.028	0.758
source	SVDB	18	−0.191	0.056
source	SWaT	1	−0.161	0.000
source	Stock	8	−0.040	0.500
source	TAO	2	−0.005	0.000
source	TODS	13	−0.004	0.385
source	UCR	70	0.010	0.514
source	WSD	20	0.006	0.450
source	YAHOO	30	0.313	0.667

Table 26: Real-data characterization, TSB-AD-M (J_p minus energy, both windowed, σ=0.35, 3 seeds).

split	bucket	n	mean gap	win-rate
type	mixed	13	0.010	0.385
type	sequence-only	167	−0.020	0.383
length/window	< 0.5	23	0.002	0.304
length/window	0.5–1	14	0.000	0.286
length/window	1–2	34	−0.031	0.353
length/window	2–5	23	−0.013	0.217
length/window	> 5	86	−0.022	0.477
avg. length	≤ 5	12	0.011	0.417
avg. length	6–25	5	−0.003	0.400
avg. length	26–100	31	−0.054	0.226
avg. length	101–400	59	−0.038	0.390
avg. length	> 400	73	0.007	0.438
source	CATSv2	5	0.047	0.600
source	CreditCard	1	0.024	1.000

Continued on next page

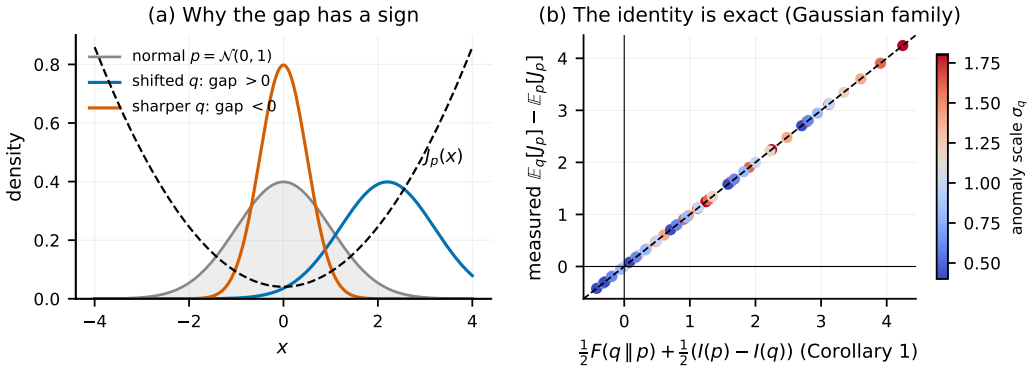


Figure 2: Anatomy of the detection gap (Corollary 1). (a) The Hyvärinen objective J_p as a parabola in a scalar deviation x: a mean-shifted anomalous q (positive gap, upper-tail detectable) versus a *sharper* q chosen with $\Delta < 0$ (equivalently $I(q) - I(p) > F(q\|p)$; e.g. $q = \mathcal{N}(0, \frac{1}{4})$ against $p = \mathcal{N}(0, 1)$), the negative-gap inversion an upper-tail test must miss. (b) The measured gap $\mathbb{E}_q[J_p] - \mathbb{E}_p[J_p]$ against the closed-form Corollary 1 value $\frac{1}{2}F(q\|p) + \frac{1}{2}(I(p) - I(q))$ across a Gaussian (μ, σ_q) family: the points lie on the identity line, confirming the identity numerically including its negative regime.

split	bucket	n	mean gap	win-rate
source	Daphnet	1	−0.020	0.000
source	Exathlon	25	0.000	0.240
source	GECCO	1	0.526	1.000
source	GHL	23	0.012	0.739
source	Genesis	1	0.066	1.000
source	LTDB	4	0.007	0.750
source	MITDB	11	0.058	0.364
source	MSL	14	0.011	0.429
source	OPPORT.	7	0.021	0.857
source	PSM	1	−0.003	0.000
source	SMAP	25	−0.056	0.280
source	SMD	20	−0.122	0.250
source	SVDB	28	−0.055	0.179
source	SWaT	2	−0.013	0.000
source	TAO	11	0.010	0.364

G FIGURES FOR THE MAIN RESULTS

This appendix collects the figures accompanying the main results: the anatomy of the signed detection gap (Figure 2), the noise-scale selection on the tuning split (Figure 3), the equal-treatment aggregation audit (Figure 4), the detection and inversion mechanism on real series (Figure 5), the J_p-versus-energy complementarity (Figure 6), the per-source breakdown (Figure 7), and conformal calibration under drift (Figure 8). Each caption cross-references the appendix table carrying the underlying numbers.

H DEEP DIVES: WHERE METHODS FAIL AND WHO FAILS TOGETHER

This appendix uses per-series VUS-PR for nineteen methods (mean over available seeds) to ask questions the aggregate leaderboard cannot answer. Sixteen methods cover all 350 series in this collection; KShapeAD (339), LSTMAD (274), and the windowed energy (349) are partial and enter only the analyses stated below. StreamVAE covers 350 here because the per-series mean unions its seeds (per-seed coverage 318/350/349 for 2024/2025/2026);

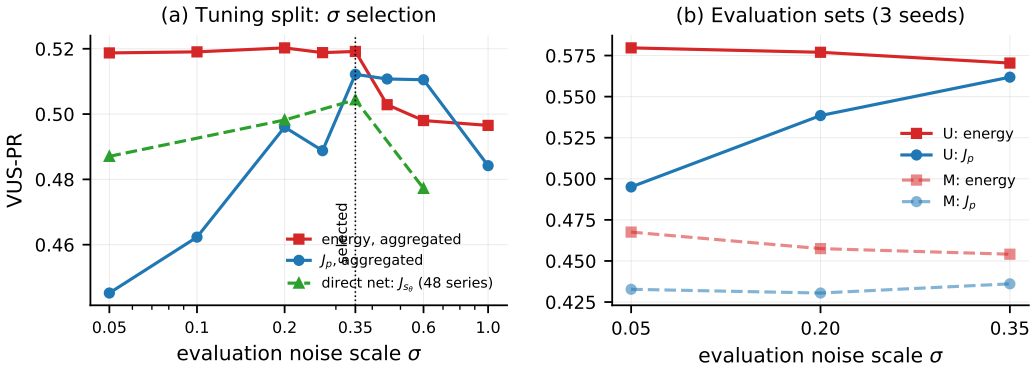


Figure 3: Noise-scale selection. (a) Disjoint tuning split (energy net: the 46 series complete under both statistics; direct net: all 48). The window-aggregated energy (the network's unnormalized negative log-density; Section 4) is nearly flat in σ (maximum at σ=0.2) while aggregated J_p rises and meets it near $\sigma = 0.35$, the value fixed for the J_p/direct-score evaluation runs; the direct score network's J_{s_θ} (which has no energy) also peaks at $\sigma = 0.35$. (b) The same comparison on the evaluation sets (3 seeds, eval-set ablation). The two statistics converge in accuracy while their median per-series rank agreement stays near 0.90 (Table 14).

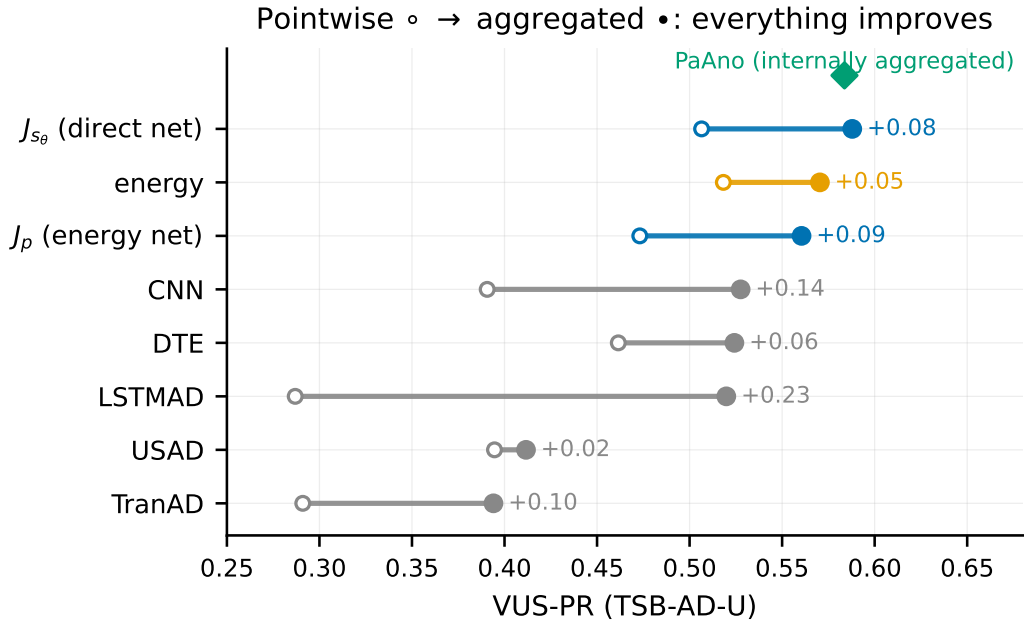


Figure 4: Equal-treatment aggregation on TSB-AD-U (VUS-PR; baseline rows use seed-2024 saved scores, our rows are 3-seed means, as in Table 3). Each baseline's own saved anomaly score is window-averaged exactly as ours; the dumbbells connect the pointwise score (left) to its aggregated value (right). Our rows are at σ=0.35 (energy: the 349 series that completed). Every method rises, and the field compresses; the score statistics stay at the top while the baseline ordering changes (LSTMAD overtakes TranAD and USAD once aggregated). PaAno is marked as already internally aggregated (patch-aggregated by construction), so it is not re-aggregated here. Tabulated in Table 3.

its main-leaderboard row reports the 318-series all-seed cohort. All panels derive from `results/tables/deepdive_long_U.csv`.

Who wins where. Figure 13 breaks mean VUS-PR down by source domain. No method dominates: across the seventeen domains with at least seven series, eleven different methods hold the

Table 27: Equal-treatment audit on the common cohort: the 246 series scored by all five saved-score baselines (seed 2024, VUS-PR). Every lift persists; CNN and LSTMAD overtake USAD once aggregated (LSTMAD already exceeds TranAD pointwise on this cohort), consistent with the per-method cohorts of Table 3. Source: `equal_treatment_U_common.csv`.

Method	point	aggregated	gain
J_{s_θ}, direct net (ours)	0.494	0.579	+0.085
CNN	0.339	0.548	+0.209
DTE	0.470	0.549	+0.079
LSTMAD	0.289	0.519	+0.229
TranAD	0.232	0.376	+0.144
USAD	0.388	0.411	+0.023

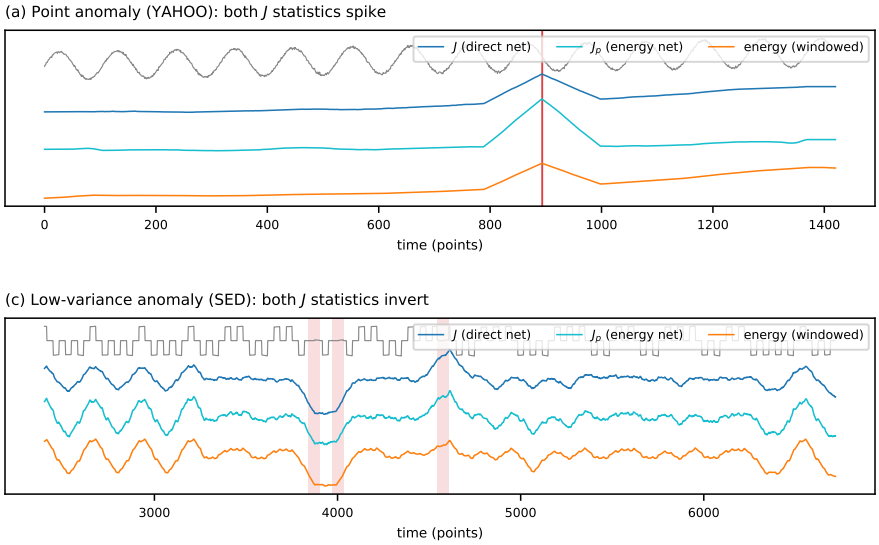


Figure 5: Detection mechanism on real series (signal gray; J of the headline *direct* net blue; J_p of the energy net cyan; windowed energy orange; anomalies red). Top: YAHOO point anomaly, where both J statistics spike. Bottom: low-variance SED anomaly where both J statistics *invert*, dropping below their normal level over the anomalous span (direct net: anomalous-span mean -676.3 vs. normal -656.3; `mechanism_sed.csv`), consistent with the sharp-q class of Corollary 1 (the exact $I(q) - I(p)$ versus $F(q\|p)$ balance was not estimated); discussed in Section 6 ("The limits").

per-domain best. The Hyvärinen statistics lead on SMD, WSD (energy-net J_p) and UCR (direct net); the windowed energy leads on IOPS; Sub-KNN leads on MSL and MITDB and KShapeAD on SVDB and LTDB (the quasi-periodic and electrocardiogram domains, where subsequence distance suits morphology matching); Sub-PCA leads on OPPORTUNITY and NEK; StreamVAE on Exathlon and MGAB; Time-RCD on YAHOO and TODS; PaAno on SMAP; AxonAD on NAB; CNN on Stock. Score statistics lead on short-anomaly telemetry domains, while distance and subspace methods lead where anomalies are long activity segments or morphology changes in periodic signals. This is the complementarity of Appendix F (Table 25) recurring at the domain level, across the whole field.

Failure complementarity. Figure 14 reports, over the nine strongest methods on their 349 common series (the windowed energy at σ=0.35 lacks one), the fraction of that cohort where the row method succeeds (VUS-PR > 0.5) while the column method fails (< 0.2). The matrix is far from symmetric and far from zero: on 20% of the cohort the direct network succeeds where StreamVAE fails (14% against AxonAD); PaAno and the direct network each succeed where the other fails on 6–8%. An oracle selecting the better of just {direct network, PaAno} per series reaches 0.703 VUS-PR; an oracle over all nineteen methods (each on its available series) reaches 0.827, and over the sixteen full-coverage methods 0.816, against 0.588 for the best single method. Most of the bench-

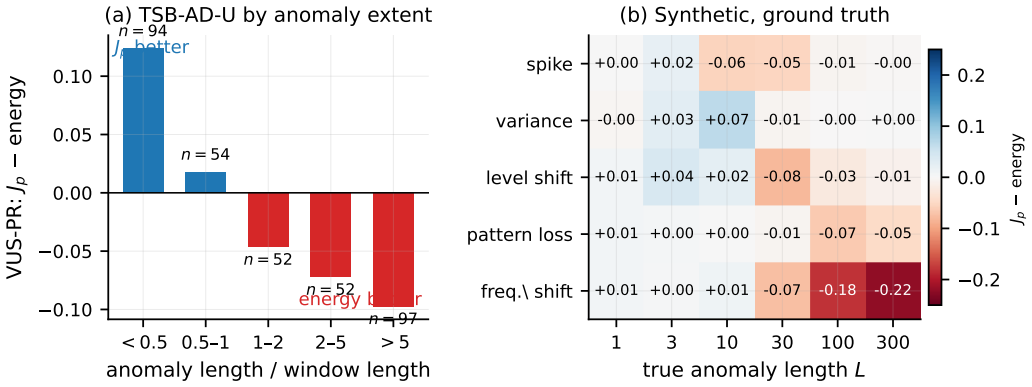


Figure 6: Complementarity of J_p and the energy on TSB-AD-U ($\sigma = 0.35$, windowed, 349 series, 3 seeds). (a) Mean J_p-minus-energy VUS-PR by anomaly-length-to-window bin (bars; per-bin n annotated): J_p leads on short anomalies (shorter than half a window $+0.124$), the energy on long ones (longer than five windows -0.097); by anomaly type, point-only series favor J_p by $+0.377$ (Table 25). (b) A synthetic heatmap over anomaly type and ground-truth length L: short anomalies ($L \leq 10$) often favor J_p (exceptions: `spike` L=10, `variance` L=1), energy ahead for long anomalies. Post-hoc characterization: the disagreement between the two statistics is organized by anomaly extent (Table 25).

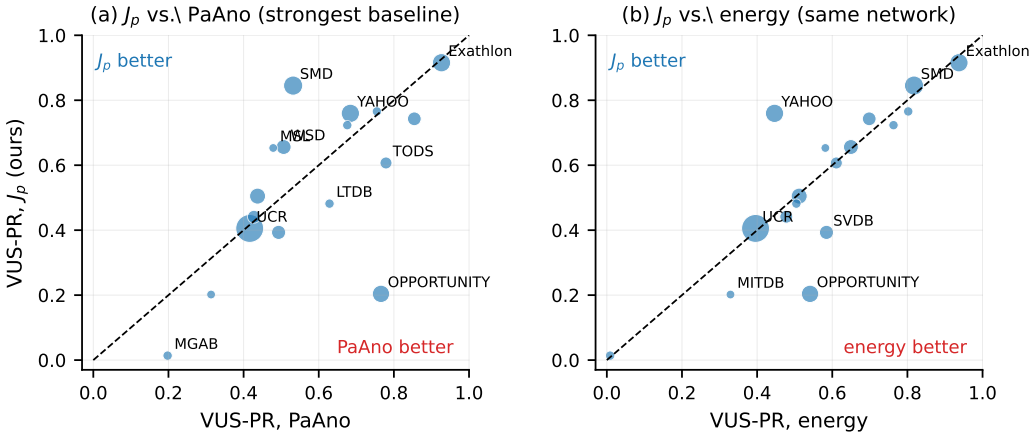


Figure 7: Per-source VUS-PR on TSB-AD-U (energy-net J_p, $\sigma = 0.35$, mean over 3 seeds; sources with ≥ 7 series, marker area $\propto n$). (a) J_p against PaAno: J_p ranks in the top three on 4 of the 7 largest sources, and the deficit to PaAno is concentrated in the long-anomaly OPPORTUNITY and ECG families. (b) J_p against the same network's windowed energy. The losses fall in the long-anomaly extent regime characterized in Table 25. The direct net's per-source ranks are given in Appendix D. Tabulated in Table 9.

mark's remaining headroom is therefore in method *selection*, not in any single detector; only 14 of 350 series (4%) are failed by every method simultaneously.

Which domains are hard. Figure 15 plots the per-series oracle (best over all methods) by domain. MGAB is the hardest domain by median (oracle median 0.36: its synthetic Mackey–Glass anomalies are partially detected by every method class but well detected by none), while MITDB (14% of series unsolved) and UCR (17% unsolved) contain the outright failures. At the other end, YAHOO, Stock, Exathlon and OPPORTUNITY are nearly saturated at the oracle level: on these domains the open problem is choosing the right method, not building a better one.

Anomaly type. Figure 16 splits performance by TSB-AD's anomaly-type metadata. On point-only series the two score statistics and Sub-KNN are the clear leaders; on sequence-only series the

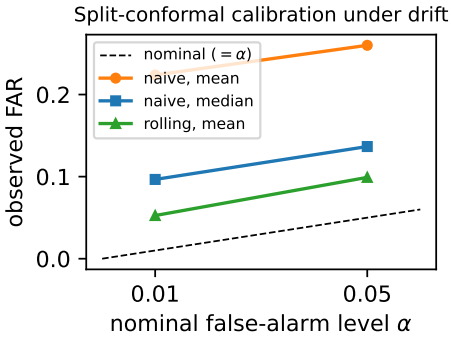


Figure 8: Naive split conformal calibration over-alarms under drift on TSB-AD-U (headline direct net, 3 seeds, exact conformal rank, final-20% calibration; `conformal_v2_U.csv`). Realized false-alarm rate against target α: at $\alpha = 0.01$ the FAR is mean $0.224 \gg$ median 0.097, and the over-alarm is proportionally worse at small α (FAR/$\alpha \approx 22$ at 0.01 vs. ≈ 5 at 0.05). This mean $\gg$ median, right-skewed shape is consistent with heterogeneous per-series additive gaps under non-exchangeability (Barber et al., 2023); label-free rolling recalibration (green) reduces the realized FAR toward nominal without a formal guarantee under arbitrary drift (Table 29).

field compresses and PaAno/Time-RCD are competitive. This extends the anomaly-extent characterization of the main text from our two statistics to the full method set.

Who fails together. Figure 17 clusters methods by the Spearman correlation of their per-series VUS-PR profiles. Three families emerge: the two Hyvärinen statistics and the energy cluster tightly (pairwise correlation ≥ 0.76, consistent with the rank agreement of Table 14); the reconstruction/forecast family (USAD, CNN, TranAD, TimesNet, LSTMAD) forms a second block; and the subsequence-distance methods (Sub-KNN, MatrixProfile, KShapeAD) a third, with PaAno and Time-RCD sitting between the last two. Ensemble value lies across blocks, not within them, consistent with the complementarity matrix.

Instances. Figures 9–12 show four series selected programmatically to maximize the performance gap between method families (signal on top, anomalies shaded; below, each method's windowed score trace with its VUS-PR on this series). (a) A SMAP sensor ramp with a brief slope glitch: the direct network's J spikes exactly at the anomaly (0.98) while all six other methods are essentially blind (≤ 0.01); the glitch barely perturbs amplitude, but it perturbs the local score field. (b) The mirror image on an OPPORTUNITY activity segment: USAD (0.99) and LSTMAD (0.93) elevate their reconstruction error across the entire sustained segment, while both score statistics fluctuate with local structure and miss the long in-support shift (0.04–0.05), a small-gap case of Corollary 1 diluted over a long segment. (c) A UCR environment series where only DTE's diffusion-time posterior fires (0.64 vs. ≤ 0.01 for all others), evidence that the diffusion-time axis carries information the score field at a single noise scale does not. (d) An unsolved UCR activity series: a subtle gait-pattern change that leaves every trace flat in the anomalous region while firing spuriously elsewhere; 14 of 350 series share this fate. These instances are the per-series texture behind the complementarity matrix: the methods fail on *different* series for *interpretable* reasons.

I Reproducibility Notes

This appendix documents the patched and failed baselines, the cite-only rows, the benchmark caveats that apply to every method, the conformal false-alarm result, the compute footprint, and the researcher-degrees-of-freedom disclosure for the σ selection.

Patched baselines. Two TSB-AD models ship with a transposed `MinMaxScaler` call, `fit(X.T)/transform(X.T).T` on the window matrix $X \in \mathbb{R}^{n_{\text{win}} \times d}$: fitting on $X^\top$ computes min/max *per window* (not per feature), and the transform crashes whenever train and test window counts differ (`OCSVM`, `AE`). Our patch replaces it with the stateless map $x_{ij} \mapsto (x_{ij} - \min_j x_{ij})/(\max_j x_{ij} - \min_j x_{ij})$, i.e. each window scaled to $[0, 1]$ by its own min/max over its d

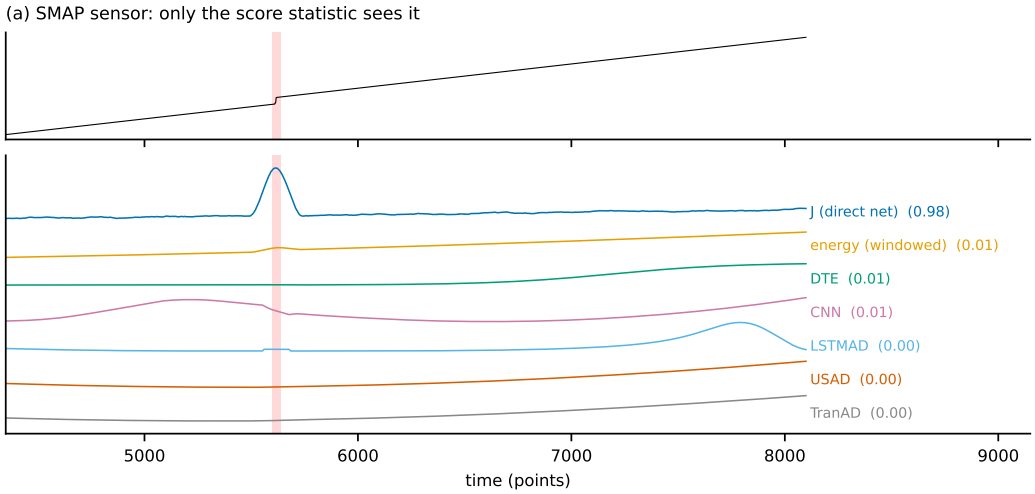


Figure 9: Case (a): SMAP sensor. Only the direct network's window-averaged J detects the slope glitch (VUS-PR in parentheses; anomaly shaded).

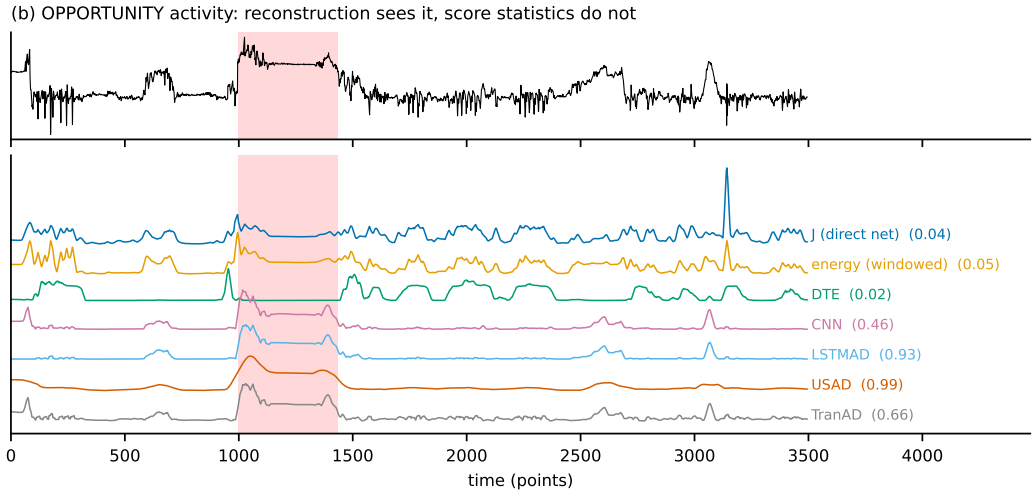


Figure 10: Case (b): OPPORTUNITY activity segment. Reconstruction detectors sustain elevated error over the long anomaly; score statistics track local structure and miss it.

lag features (reduction axis j). This reproduces exactly what the shipped code computed when fit and transform arrays coincide (its train-time behavior) and extends it statelessly to test windows; it is *not* ordinary per-feature scaling fitted on training samples, and no fitted statistics are carried across windows. No other behavior changes; the patch is documented in the supplementary material and applies to TSB-AD commit `6beac72e`, to which the editable install is pinned.

Failed baselines (documented, not silently dropped). Several methods fail inside TSB-AD's own code on portions of the benchmark and are reported with reduced coverage or cited from the bundled leaderboard CSV rather than re-run:

- **Donut** and **LSTMAD**: raise "stack expects a non-empty TensorList" on short training prefixes: Donut fails on all series (U: $0/350$, M: $0/180$ across all seeds), LSTMAD on the short-prefix subset (U: $274/350$ ran, i.e. 76 failed; M: $147/180$ ran, i.e. 33 failed per seed). Over three seeds this gives, per method and suite: Donut $350 \times 3 = 1050$ (U) and $180 \times 3 = 540$ (M) failed series-runs; LSTMAD $76 \times 3 = 228$ (U) and $33 \times 3 = 99$ (M).

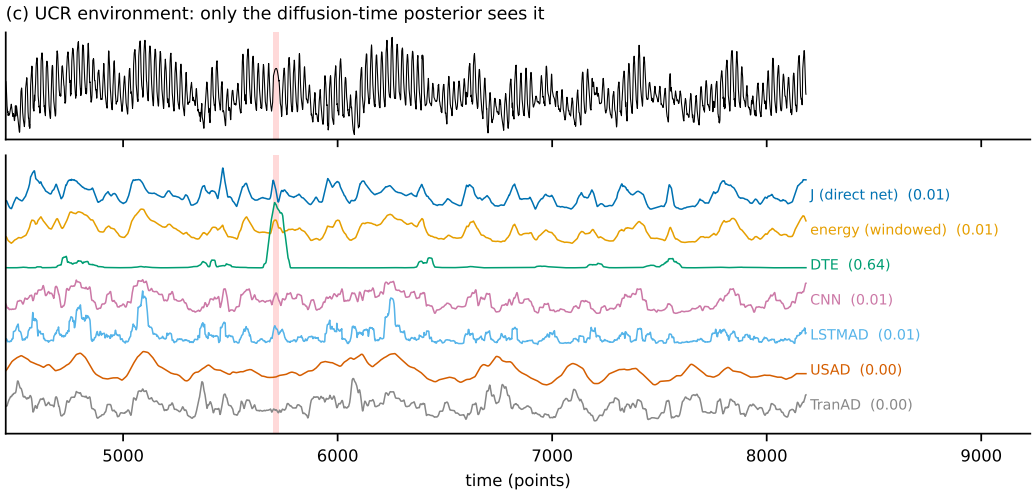


Figure 11: Case (c): UCR environment series. Only the diffusion-time posterior (DTE) fires at the anomaly.

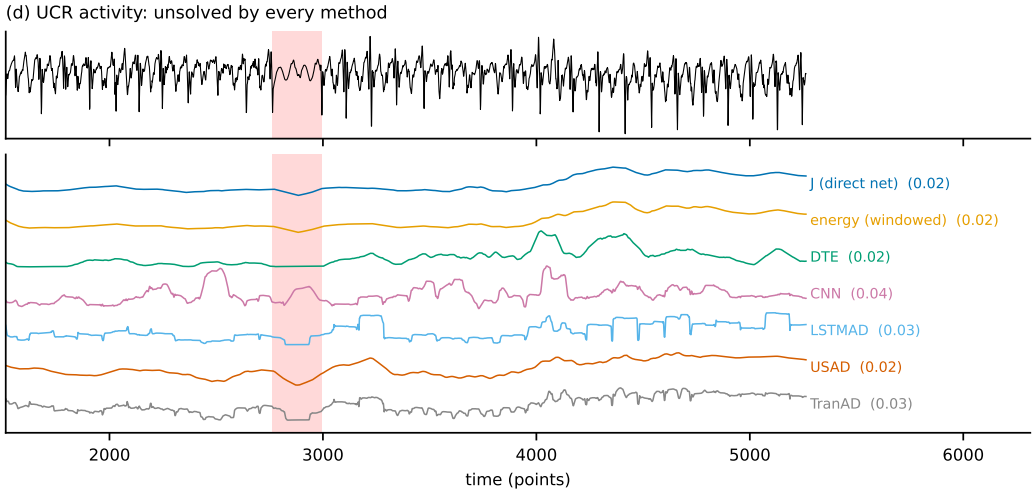


Figure 12: Case (d): unsolved UCR activity series; every method is quiet in the anomalous region and fires spuriously elsewhere.

- **SAND** and **KShapeAD**: sampling failures on short series (SAND ran 191/350 on U; KShapeAD 339/350).
- **Time-RCD (M)**: checkpoint strict-load failure on TSB-AD-M (the released checkpoint does not match the M input shape); we report Time-RCD on U only.
- **MMPAD**: the U run was stopped after 10 series (none failed) and not resumed; on M, 155 series were attempted before the job ended, 16 of which failed inside MMPAD's own code (139/180 ok). Single seed, reported with its coverage and excluded from the main-text tables.
- **xLSTMAD**: CUDA kernel build failure in our environment (0 series ran on U and M); its score is cited from the bundled leaderboard CSV (U VUS-PR 0.398, Liu & Paparrizos, 2026; method: Faber et al., 2025).
- **AxonAD** (TSB-AD-M): 18 of the 180 M series fail inside AxonAD's own code, so its full attempted-coverage number (Table 1) is computed over the 162/180 series that complete under seed 2024; the resulting 0.494 matches the published 0.495. A later seed-2025 re-run was interrupted after 121 series (120 ok, 0.487; on the 120 common series the two seeds give 0.480/0.487); it is not pooled (`axonad_M.csv`).

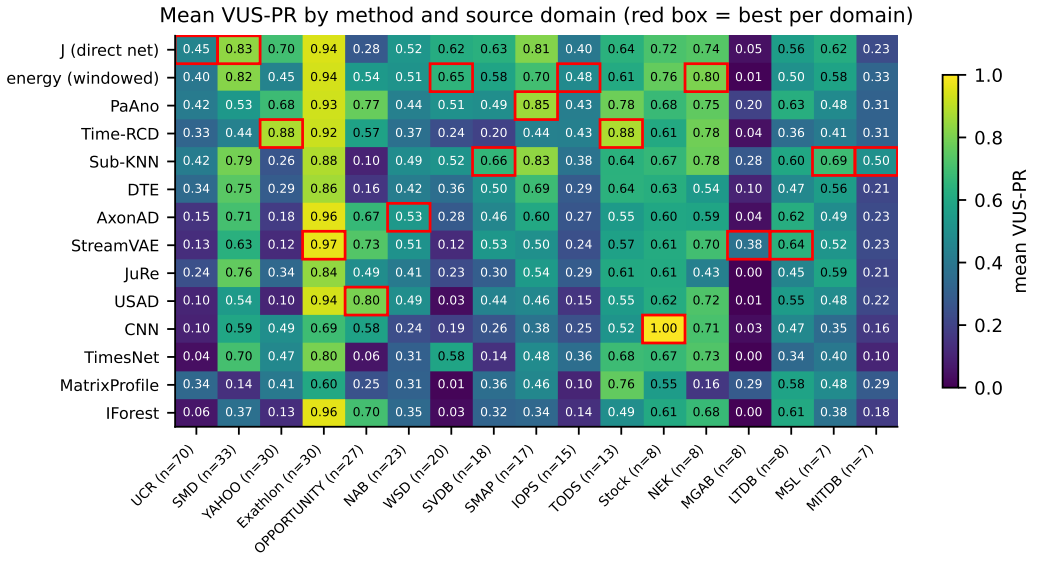


Figure 13: Mean VUS-PR by method and source domain (TSB-AD-U, domains with ≥ 7 series; red box marks the per-domain best). No single method dominates; score statistics lead on point-anomaly domains, distance and reconstruction methods on long-anomaly and ECG domains.

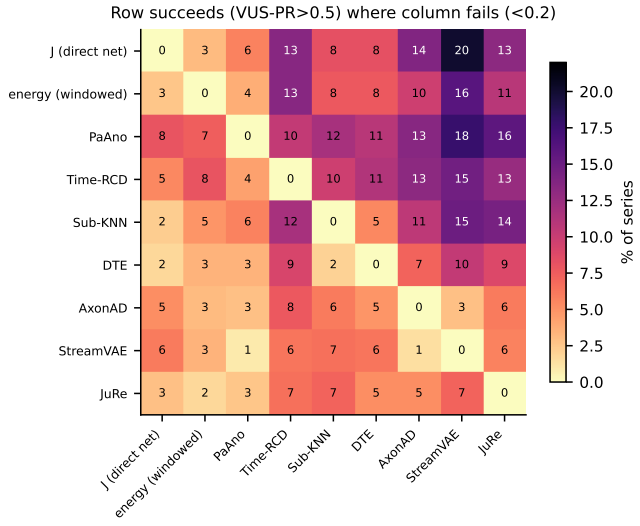


Figure 14: Failure complementarity, nine strongest methods on their 349 common series: percentage of that cohort where the row method attains VUS-PR > 0.5 while the column method scores < 0.2.

Reduced-coverage rows in Tables 6–8 carry their true n_{series} so the reader can see the coverage.

Cite-only rows. The following TSB-AD-U VUS-PR values are taken from the bundled TSB-AD leaderboard CSVs (Liu & Paparrizos, 2026; identical metric and 350-series split) and are not re-run: Time-RCD (MAFT) 0.586 (Lan et al., 2026), TSPulse fine-tuned/zero-shot 0.552/0.480 (Ekambaram et al., 2026), CHARM 0.501 (Dutta et al., 2025), xLSTMAD 0.398 (Faber et al., 2025), MOMENT fine-tuned/zero-shot 0.386/0.379 (Goswami et al., 2024), TimesFM 0.300 (Das et al., 2024), Lag-Llama 0.273 (Rasul et al., 2023), Chronos 0.272 (Ansari et al., 2024), and OFA 0.237 (Zhou et al., 2023). Per-file provenance: `cite_only.csv`.

TSB-AD caveats (identical for all methods). No labels reach any detector; normalization and calibration use the training prefix only. The tuning split is disjoint from both evaluation suites

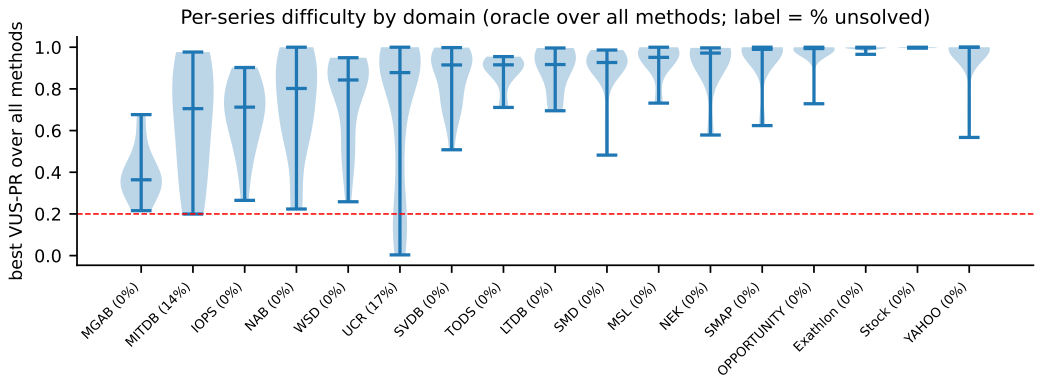


Figure 15: Per-series difficulty by domain: distribution of the oracle VUS-PR (best over all nineteen methods available per series); the axis label reports the fraction of series unsolved (oracle < 0.2).

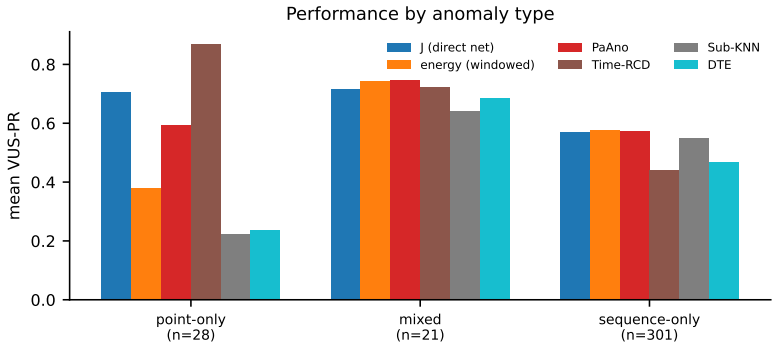


Figure 16: Mean VUS-PR by TSB-AD anomaly-type label for six leading methods.

(verified). A minority of training prefixes contain anomalies ($35/350$ on U, $20/180$ on M), and the whole series, prefix included, is scored; this is a property of the benchmark and applies uniformly to every method compared. The RANDOM row (VUS-PR $\approx$ anomaly rate) validates the metric choice and appears in the leaderboard and audit tables.

Conformal false-alarm rates under an alternative configuration. Table 28 reports a conformal run with a different configuration (energy net, 50% calibration split, interpolated quantile, seed 2024), shown for comparison; the procedure used in the paper is Table 29 (direct net, final-20% calibration, exact rank, 3 seeds). The qualitative finding is identical in both: split-conformal calibration over-alarms under drift because prefix and remainder are not exchangeable.

Table 28: Alternative conformal configuration (energy net, 50% split, interpolated quantile, seed 2024; `conformal_fix_U.csv`), for comparison with the procedure of Table 29. Nominal level α vs. empirical FAR; the guarantee is void under drift.

α	FAR mean	FAR median	frac series FAR$\leq \alpha$	TPR mean	n
0.01	0.1805	0.0642	0.1629	0.6706	350
0.05	0.2205	0.1098	0.2057	0.7169	350

Diagnosis and label-free fix. The over-alarm is consistent with heterogeneous per-series additive gaps under non-exchangeability (Barber et al., 2023): realized FAR is centered above α, right-skewed (mean $\gg$ median), and proportionally worst at small α. Under the procedure of Table 29 the naive FAR is mean 0.224 / median 0.097 at α=0.01 and mean 0.260 at α=0.05. Label-free rolling recalibration on a trailing, mostly-normal window reduces the realized false-alarm rate toward nominal without a formal guarantee under arbitrary drift, trading detection for calibration (Table 29).

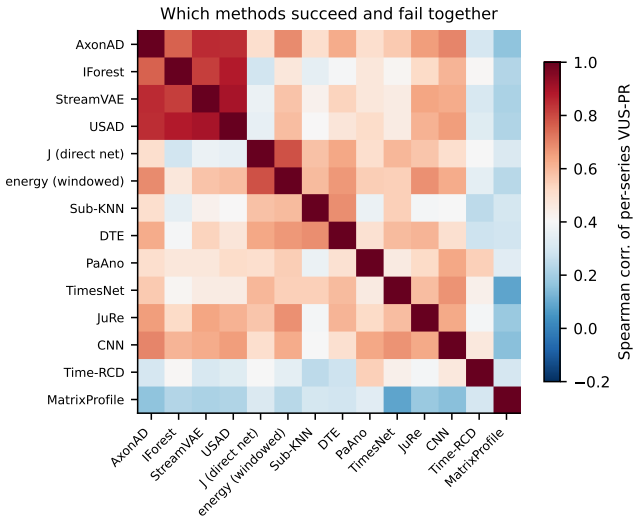


Figure 17: Method similarity: Spearman correlation of per-series VUS-PR profiles, hierarchically ordered. Score-based, reconstruction-based, and subsequence-distance families fail together within blocks.

Adaptive conformal inference (ACI) methods that consume realized-error feedback are inadmissible in this unsupervised one-class setting, where the "false alarm?" label is never observed.

Table 29: Naive vs. label-free rolling recalibration on TSB-AD-U, headline direct net (350 series, 3 seeds, final-20% calibration, exact conformal rank; `conformal_v2_U.csv`). Rolling recalibration cuts the mean false-alarm rate toward the nominal level at a cost in true-positive rate, without a formal guarantee under arbitrary drift.

α	method	FAR mean	FAR median	TPR mean
0.01	naive	0.224	0.097	0.733
0.01	rolling	0.053	0.030	0.391
0.05	naive	0.260	0.137	0.765
0.05	rolling	0.099	0.067	0.527

Runtime. Wall-clock per series (fit + score, seed 2024, U suite; `runtime_U.csv`). All neural methods below ran under the identical runner on the same 8-GPU node (one GPU worker per job), so their times are directly comparable: the direct score net averages 3.0 s per series (17 minutes for the full 350-series suite), including exact/Hutchinson divergence evaluation; PaAno 10.1 s; the energy net 9.5 s; CNN 9.7 s; LSTMAD 24.3 s; TranAD 26.1 s; DTE 28.8 s. The classical baselines ran on CPU instances and are listed for scale only (IForest 2.2 s, Sub-KNN 114.8 s); means over completed series, medians and totals in the CSV.

Compute and scale. Experiments ran on 8× L40S GPUs locally plus 7 cloud processing jobs for a subset of baselines (results synced per `results/MANIFEST.md`). The manifest records 73,254 result files (69,371 with status ok) across 322 (group, suite, method, seed) cells. Every shipped script is fingerprinted by `sha256/8` in `results/MANIFEST.md`.

Researcher-degrees-of-freedom disclosure (σ selection). We disclose verbatim: *σ was selected on the disjoint TSB-AD tuning split (48 series); the complete sweep is in Table 11. The decision to sweep σ at all was prompted by an evaluation-set observation at σ=0.05 (where windowed J_p loses clearly to the windowed energy); an evaluation-set σ sweep therefore exists as an ablation, and we report it as such rather than presenting the tuning choice as fully pre-registered.* The headline σ=0.35 is the raw-J_p maximizer on the tuning split (0.0015 above the runner-up σ=0.45); the fair-criterion maximizer is σ=0.6, whose J_p-minus-energy gap is 0.0195 higher (+0.0125 vs. −0.0070), so the headline choice is, if anything, conservative for the J_p-vs-energy comparison. The windowed

energy is nearly flat in σ on the tuning split and peaks at σ=0.2 (0.5203), the scale used for the energy rows of the main leaderboards.