PRINCIPAL-CURVE NOVELTY GEOMETRY: FALSIFIABLE ATTRIBUTION OF NOVELTY IN FROZEN SEQUENTIAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc novelty methods usually reduce representation unusualness to a single scalar, obscuring whether an observation is unusual because of its position along a learned trajectory, its deviation from that trajectory, or its rate of progression. We introduce *principal-curve novelty geometry* (PCNG), a post-hoc measurement contract for frozen sequential representations that asks which geometric interpretation is supported and when attribution should be refused. PCNG fits a training-only principal curve, tests whether the resulting progression geometry is admissible, and, when permitted, derives progression, transverse deviation and traversal rate on a common rarity scale calibrated against the training reference. Targeted interventions test these interpretations, while unit-preserving randomisation evaluates residual coupling beyond the dependence already present within trajectories. Across 117 runs in degradation prognostics spanning 13 frozen configurations, failure-mode interventions support the transverse interpretation in 35 of 36 directed tests, whereas the rate interpretation is unsupported. Cross-unit scrambling reduces the transverse hit rate from 0.972 to 0.667, while a within-unit control remains at 1.000. Residual coupling is heterogeneous, precluding an independence claim; the admissibility gate also varies across encoders and rejects some external-domain runs. Operationally, transverse rarity yields a 4.9% false-alarm rate versus 29.2% for a learned early-degradation health monitor at matched coverage, while distance-based novelty remains competitive. PCNG therefore contributes falsifiable attribution rather than universally better detection.

1 INTRODUCTION

From novelty detection to novelty attribution. A reliability score usually answers one question: *how unusual is this example?* For a sequential representation, however, that answer is incomplete. A unit can be unusual because it has moved away from the path traced by the reference population, because it has progressed beyond the part of that path previously observed, or because it traverses a familiar path at an unfamiliar rate—cases that can call for different responses even when the scalar score is identical. We therefore study *representation-novelty attribution*: given a frozen sequential representation and a reference population, can we decompose novelty into geometrically interpretable modes and know when those interpretations are supported?

Downstream prediction error does not supply this distinction: on N-CMAPSS, seven readouts spanning two to thirteen parameters fall within $1.19\times$ of one another, and twenty-six representations fall within $1.1\times$ across two bearing datasets. A metric that flat is a poor instrument for comparing representation structure. Nor is novelty predictive uncertainty—Appendix U tests whether transverse novelty identifies high-error windows and finds no separation from random selection. Our object is the geometry of a frozen representation, not a calibrated estimate of prediction error.

PCNG is a measurement contract, not three scores. On frozen embeddings, we fit a post-hoc principal curve (Hastie & Stuetzle, 1989), which provides an arc-length parameterization and nearest-point projection. PCNG then defines three raw coordinates from that geometry and sequence order: normalized projected arc length s, transverse distance d, and temporal progression rate $\dot{s}$. We map them to reference-calibrated rarities $(\mathcal{U}_{\parallel}, \mathcal{U}_{\perp}, \mathcal{U}_{\dot{s}})$ on a common dimensionless scale. The primitives

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.

are familiar; the contribution is the contract around them. First, an admissibility check asks whether one progression curve is an appropriate object before any semantic reading is allowed. Second, the coordinate names are treated as hypotheses rather than truths: targeted interventions ask whether the coordinate responds to the deficiency it names. Third, residual statistical coupling is a separate question, evaluated with a unit-preserving reference that respects trajectory structure. Finally, the readout is allowed to return a negative result or refuse the dataset. The governing principle is simple: *computability is not validity, and validity is not independence.*

The separation matters because the three can come apart, as they do here (Table 1). PCNG therefore does not turn geometry into semantics by naming axes after the fact; it makes the semantic license itself an empirical output of the protocol.

Why principal-curve geometry. For a frozen sequential representation with a dominant progression direction, a principal curve provides the path from which PCNG derives the decomposition used for attribution: position *along* the population trajectory, deviation *away* from it, and rate of motion *through* it. PCNG is tied to no encoder objective and updates no encoder parameter. Degradation prognostics is our testbed because progression is physically meaningful there, not because the construction is degradation-specific. The PCNG admissibility gate makes that scope explicit rather than assumed (Section 7).

Contributions.

1. **Representation-novelty attribution as a post-hoc measurement problem.** We formulate the task of attributing why a frozen sequential representation is novel and instantiate it with PCNG: an admissibility gate, a principal-curve coordinate frame, and reference-calibrated rarity, all without retraining the encoder.
2. **Falsifiable semantics and unit-aware residual-coupling diagnostics.** Coordinate interpretations are licensed by targeted interventions rather than by their construction alone, while a unit-preserving randomisation asks whether residual coupling exceeds the dependence already present inside real trajectories. A cross-subset failure-mode shift supports the transverse interpretation in $35/36$ directed cells; the operating-condition intervention selects $\mathcal{U}_{\hat{s}}$ in only $21/36$, so we decline the rate interpretation; and $4/9$ in-domain runs of our MonoCNN encoder exceed their own unit-preserving 97.5th-percentile threshold $p_{97.5}$, so we make no independence claim. More broadly, the audit shows that window-level sample size can miscalibrate evidence about representation geometry: on PRONOSTIA, an apparent $7.1\times$ residual-coupling ratio under an independent-window reference becomes $1.17\times$ under a unit-preserving reference.
3. **Operational validation with explicit refusal.** The licensed transverse component survives residual-scramble ($0.972 \to 0.667$, with a within-unit control at 1.000) and reference-resampling checks, and supports a near-failure alarm at a 0.049 false-alarm rate against 0.292 for the learned health control. The protocol also reports where the claim stops: external-domain admissibility is mixed under the same post-hoc progression geometry, and classical distance-based novelty remains competitive.

2 Related Work

Appendix A gives the full treatment of each neighbouring literature; we state the essential positioning here.

Principal curves as coordinates, not conclusions. Principal curves give a self-consistent one-dimensional summary of a distribution (Hastie & Stuetzle, 1989), with a long literature on making the estimator well posed (Tibshirani, 1992; Kégl et al., 2000; Ozertem & Erdogmus, 2011). PCNG uses that mature object not as a dimensionality reduction algorithm but as a coordinate system whose along-curve and transverse quantities become *candidate measurements* that the geometry alone does not license.

Scalar novelty and uncertainty answer a different question. Mahalanobis distance (Mahalanobis, 1936; Lee et al., 2018), k-nearest-neighbour (kNN) novelty (Sun et al., 2022) and kernel density scores (Parzen, 1962) return a scalar measure of unusualness, and ensembles, dropout and conformal

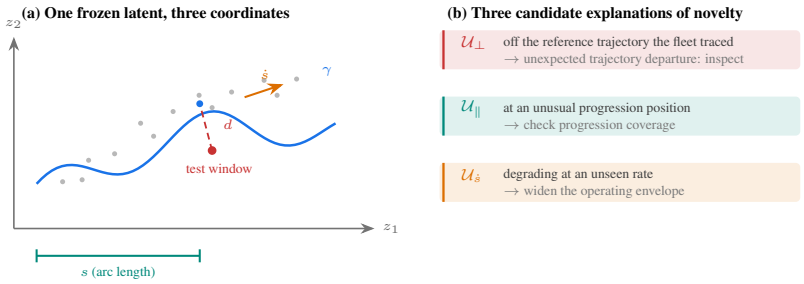

Figure 1: **From scalar novelty to candidate geometric attributions.** *(a)* A principal curve fitted post hoc to a frozen encoder provides a nearest-point projection; PCNG derives normalized projected arc length s, transverse distance d, and temporal progression rate $\dot{s}$ from that geometry and sequence order. *(b)* After reference calibration, these coordinates propose three candidate explanations for why a representation is unusual. PCNG does not accept those labels by construction: their semantic interpretations are hypotheses validated by intervention, and one of the three fails in our experiments.

methods return scalar or interval statements about predictive reliability (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Vovk et al., 2005; Angelopoulos & Bates, 2023). PCNG replaces neither family and claims no superior out-of-distribution (OOD) performance. Its target is orthogonal—attributing unusualness to movement along a learned trajectory, deviation away from it, or rate through it—which is why matching Mahalanobis distance in Section 8.2 is compatible with the contribution.

Measurement validity is the closest conceptual precedent. Machine-learning quantities increasingly function as measurements of latent constructs, carrying the obligation to establish what they actually measure rather than infer validity from a convenient proxy (Jacobs & Wallach, 2021; Raji et al., 2021; Liao et al., 2021; Hutchinson et al., 2022). PCNG operationalises that obligation for representation geometry: a transverse distance is not a failure-mode novelty merely because it is transverse, but only if an intervention selects it, a scramble attacks the structure carrying it, and resampling leaves the decision stable.

Frozen representations and prognostic geometry. Eleven of the thirteen frozen configurations PCNG reads come unmodified from the self-supervised time-series literature (Appendix A.6); two are our own. Closest in intent are methods building prognostic geometry into training—a polar (θ, r) representation on C-MAPSS (Thil et al., 2026; Saxena & Goebel, 2008), architectural monotonicity (Zhao & Fink, 2025), and RaPP-derived structure from autoencoder representations (Thil et al., 2025). PCNG instead freezes whatever representation it is given and asks what that representation can support post hoc.

3 PCNG: A Falsifiable Measurement Contract

PCNG has four checkpoints: freeze the representation; fit a training-only principal curve and test whether the resulting progression geometry is admissible; extract and calibrate candidate coordinates when scoring is permitted; and validate their interpretations before attaching an operational readout. A coordinate may exist numerically even when its interpretation is refused, and that is a designed output, not a hidden failure mode.

What the contract returns. PCNG separates the numerical readout from its status, returning an admissibility verdict; when scoring is permitted, calibrated rarities U_c; a semantic status for each candidate interpretation (licensed, unresolved, or refused); and a separate statement about whether any independence claim is licensed.

We report the coefficient of variation of per-unit traversal length,

$$L_u = \text{traversed arc length of training unit } u, \qquad CV(L) = \frac{\mathrm{sd}_u(L_u)}{\mathrm{mean}_u(L_u)}, \tag{1}$$

Table 1: **Computability is not validity.** Three conclusions available from numerical outputs alone but not licensed by the PCNG measurement contract. Full protocols and diagnostics are in Appendices F, E, and M.

Computable observation	Tempting conclusion	Contract verdict
$\mathcal{U}_{\dot{s}}$ is available numerically	Rate novelty is meaningful	**Refuse semantic claim:** the intended rate intervention selects $\mathcal{U}_{\dot{s}}$ in only $21/36$ cells.
Small pooled conditional $\max \lvert\rho\rvert$	The coordinates are independent	**Refuse statistical claim:** $4/9$ in-domain unit-aware runs exceed their own $p_{97.5}$; residual coupling is heterogeneous across runs.
Capacitor has 8 training units and a computable admissibility diagnostic	Treat admissibility as a categorical dataset property	**Report mixed verdicts:** post-hoc CV(L) is $0.377/0.190/0.368$, giving NO-GO/MARGINAL/NO-GO across seeds.

as the primary continuous admissibility diagnostic; GO/MARGINAL/NO-GO (CV$(L) < 0.15$, 0.15–0.35, > 0.35) are thresholded summaries of it, not an assertion that admissibility is categorical. Under the protocol, GO permits attribution, MARGINAL permits attribution with the continuous diagnostic disclosed, and NO-GO stops semantic attribution. The labels are demonstrably brittle to the threshold, to the seed, and to which progression coordinate supplies the path length (Appendices H and L), so every verdict is reported with its CV(L).

What the contract refuses. Table 1 records three findings from Sections 5.3–7: a computable quantity, a tempting conclusion, and the contract's verdict.

3.1 Stage 1: The Frozen Representation

Stage 1 supplies a frozen representation; it is not the contribution. Everything that follows is post-hoc and updates no encoder parameter. We include our MonoCNN and MonoTR plus eleven frozen literature configurations (Section 4). No remaining useful life (RUL) or failure label enters training; our temporal monotonicity term means the two encoders we train carry a monotone-degradation prior. Architectures, losses and hyperparameters are in Appendix C.

3.2 Stage 2: Geometric Attribution on a Frozen Representation

Stage 2 is label-free and post-hoc. On the frozen embeddings $z_i \in \mathbb{R}^d$ we fit a principal curve γ following Hastie & Stuetzle (1989), with a soft temporal-order regularizer (Appendix B), and project

$$\hat{z}_i = \mathrm{proj}_\gamma(z_i). \tag{2}$$

From the fitted curve, its projection, and within-unit time order, PCNG defines three coordinates on every window,

$$s_i = \frac{\ell(\hat{z}_i)}{L_\gamma} \in [0,1], \qquad \text{position along the learned path,} \tag{3}$$

$$d_i = \|z_i - \hat{z}_i\|_2, \qquad \text{transverse deviation from it,} \tag{4}$$

$$\dot{s}_i = \left.\frac{\mathrm{d}s}{\mathrm{d}t}\right|_i, \qquad \text{rate of advance along it,} \tag{5}$$

with $\ell(\hat{z}_i)$ the arc length to $\hat{z}_i$ and L_γ the total, and $\dot{s}$ computed by finite differences within a unit in time order. The curve is fitted on training data only. The health coordinate h is learned in Stage 1, whereas $(s, d, \dot{s})$ are derived after freezing, so the curve does not create h; this is what lets us ask which coordinate a task prefers.

From coordinates to dimensionless rarity. Each coordinate has a raw score and then a reference-calibrated rarity. With $V_\parallel = \mathrm{Var}(s_{\mathrm{train}})$ a fixed positive scale computed on the reference set, the raw scores are

$$R_\perp = d^2/V_\parallel, \qquad R_\parallel = s, \qquad R_{\dot{s}} = \dot{s}, \tag{6}$$

and each maps to a capped tail rarity

$$U_c(r) = -\log \max\left(q_c(r), 1/N_R\right), \tag{7}$$

where N_R is the number of reference windows and q_c is the upper-tail probability for $R_\perp$ and the two-sided tail probability for s and $\dot{s}$, both under the empirical cumulative distribution function (CDF) of the raw score on the training reference. The transverse score uses an upper tail because $R_\perp \geq 0$ measures departure magnitude; s and $\dot{s}$ use two-sided tail rarity, so $\mathcal{U}_\parallel$ should be read as unusual position rather than a one-sided late-life score. The three rarities thus share one dimensionless scale while their geometric meanings remain in R_c, making the geometric definition and its statistical calibration distinct design choices. All calibration is fitted on the reference set only and no failure label enters; Appendix B gives the tail definitions, the role of the $1/N_R$ floor, and why every reported quantity is invariant to $V_\parallel$.

Candidate semantics are hypotheses, not information dimensions. Since $R_\parallel = s$, $\mathcal{U}_\parallel$ is the rarity calibration of the progression coordinate rather than an independent dimension; because the transform is a two-sided fold rather than a monotone map, its rank correlations are not restatements of s's. Section 5.2 draws out what this implies for conditioning. $\mathcal{U}_\perp$ and $\mathcal{U}_{\dot{s}}$ differ in kind: d is transverse to the curve and $\dot{s}$ depends on the time index within a unit, so two windows at the same s in different units differ in both.

What they are taken to mean: $\mathcal{U}_\perp$ rises when a unit leaves the reference trajectory the fleet traced; $\mathcal{U}_\parallel$ marks an unusual position along progression; $\mathcal{U}_{\dot{s}}$ rises when a unit advances at a rate the reference never exhibited, so the path is familiar but the operating envelope is not. These are hypotheses the geometry suggests, not assumptions PCNG requires. Section 5.3 tests them separately, and the outcome is uneven: $\mathcal{U}_\perp$ is construct-supported by a cross-subset failure-mode shift; $\mathcal{U}_\parallel$ passes a calibration-specificity check that is close to definitional, since the ablation removes exactly the range of s occupied by the held-out windows; and $\mathcal{U}_{\dot{s}}$ is not construct-validated. Transverse attribution is therefore the licensed reading, progression serves as a calibrated positional coordinate rather than an independently validated construct, and rate remains computable but semantically unlicensed on real data. We keep all three in the frame because which coordinates survive validation is a result about this representation, not a definition of the method.

Relation to the operational statistic $\mathcal{U}_{\text{geo}}$. The operational refinement is

$$\mathcal{U}_{\text{geo}}(z_i) = \bar{d}_i \, (1 + \alpha \bar{\rho}_i) \,, \tag{8}$$

where $\rho_i = -\log \hat{p}_{\text{kNN}}(z_i)$ is a kNN local-novelty term on the full embedding (Loftsgaarden & Quesenberry, 1965) and $\bar{d}, \bar{\rho}$ are normalised by training 99th percentiles; we fix α=1.0 throughout. It is an optional density-weighted refinement of transverse distance, not a PCNG component. $U_\perp$ is the primary near-failure alarm signal; $\mathcal{U}_{\text{geo}}$ is reported as a secondary ablation.

Stage separation. The Stage 1 encoder E_θ and health head g_ϕ are frozen before any of this; the curve, reference CDFs, novelty scores and readouts are fitted without updating them. Stage 1 learns a reusable latent space, Stage 2 interrogates it, Stage 3 reads it.

3.3 Stage 3: Task Readouts

All deployment thresholds come from training data and labels score alarms only afterwards. For onset detection, we compare a cumulative-sum (CUSUM) detector with a causal rolling spike rate. For the near-failure alarm, we threshold a rolling maximum of $U_\perp$ (W=20) at the 98th percentile of the training pre-onset distribution (p_{98}) and report coverage, false-alarm rate and first-alarm position; baselines receive identical machinery (Appendix D).

4 Experimental Setup

Datasets. N-CMAPSS (Chao et al., 2021): DS01–DS03 form the development pool, while DS04 and the cross-subset split are held-out evaluations; all use the benchmark's predefined unit-held-out partitions (Appendix D). NASA Battery (Saha & Goebel, 2007), NASA Capacitor (Celaya et al., 2011), PRONOSTIA (Nectoux et al., 2012) and XJTU-SY (Wang et al., 2020) are used for the broader external readout checks.

Encoders. We evaluate 13 frozen encoder configurations—eight architectures at latent dimensions 2–320—under identical post-hoc extraction, calibration and readouts. Two architectures are ours (MonoCNN, MonoTR); six are literature baselines: TS2Vec (Yue et al., 2022), SoftCLT-TR and SoftCLT-TS2Vec (Lee et al., 2024), STEP (Thil et al., 2026), a Transformer autoencoder (Tr.AE) and a bidirectional LSTM autoencoder (BiLSTM-AE) (Hinton & Salakhutdinov, 2006; Hochreiter & Schmidhuber, 1997); SimCLR-TS (Chen et al., 2020) enters the onset sweep only. Seeds 42–44, means over three seeds unless stated.

5 Validating the Measurement Contract

The measurement contract separates two questions. *Residual coupling* asks whether the coordinates share structure beyond the progression signal and the dependence already present inside units. *Construct specificity* asks whether a named intervention preferentially moves the coordinate it is supposed to affect. Neither substitutes for the other, so we test them separately: the unit-aware diagnostic reveals heterogeneous residual coupling, and the intervention test licenses the transverse interpretation and declines the rate interpretation.

5.1 A unit-aware diagnostic of residual coupling

A naive correlation between two of the components is not evidence either way, because both depend on position along the curve. A unit late in life is simultaneously further along s, typically further from the training curve in d, and often moving faster in $\dot{s}$. Correlating the raw scores therefore measures how much each tracks ageing, not whether they track anything *else*.

We remove that shared dependence before testing. Each component is rank-transformed and residualised against s by subtracting its mean within quantile bins of s; we call the resulting statistic the *conditional rank-residual correlation*, not classical partial Spearman. We report the largest Spearman correlation among the three residual pairs as a residual-coupling summary: larger values mean more shared residual structure. Pooled over the 117 encoder–dataset–seed runs, the raw maximum pairwise correlation is 0.259 ± 0.118 and falls to 0.096 ± 0.054 after residualising on s.

The reference must preserve the unit. An independent-row reference gives a maximum-over-three-pairs scale of 0.032 at n=1726, but windows within a unit are not exchangeable. We instead circularly shift each coordinate within each unit, preserving its marginal, temporal dependence and unit offset while breaking cross-coordinate alignment. Across the 15 audited runs the unit-preserving reference is 1.2–6.5$\times$ wider than the matched independent-row reference. Four of nine N-CMAPSS runs and none of six external runs exceed their own $p_{97.5}$ (Appendix E). The effective unit is the trajectory, not the window.

5.2 What the residual-coupling diagnostic licenses—and refuses

Those residual correlations (Table 18) are small in absolute magnitude, but the unit-aware reference shows the remaining structure is not uniformly negligible: $4/9$ N-CMAPSS MonoCNN runs exceed their own $p_{97.5}$, and across the nine runs the pooled observed-to-unit-reference mean ratio is 2.23$\times$ (2.80$\times$ among the exceedances). These exceedances are positive evidence of residual coupling in those runs, not evidence that the coordinates are statistically independent. The nine jobs are three benchmark subsets by three seeds rather than nine independent fleet replications, so we do not aggregate the exceedance count into a fleet-level significance statement.

What the diagnostic refuses is the independent-row reading: treating windows as exchangeable narrows the reference and can turn ordinary within-trajectory dependence into an apparently large ratio. A retuned synthetic power check (Appendix E) detects an injected coupling of $\rho = 0.20$ in 92% of N-CMAPSS-like fleets, so the five non-exceeding runs cannot be dismissed as the output of a powerless test. We nevertheless stop short of an upper bound or independence claim because the empirical jobs are not independent fleet replicates and the power model is stylised.

A structural limit on conditioning. One fact about this decomposition limits what conditioning can show and generalises beyond our data: $\mathcal{U}_{\|}$ is a deterministic function of s, the very variable the analysis

conditions on. In a progression-relative frame, a rarity derived deterministically from progression cannot become independent of the conditioning variable in the sense a genuinely transverse coordinate can, so conditioning on s makes any pair containing $\mathcal{U}_{\parallel}$ an intrinsically weak probe. This is a property of the class of progression-relative decompositions, not of our estimator. It does not predict the observed ordering of the pairwise means ($0.070/0.057/0.035$ for $\mathcal{U}_{\perp}$–$\mathcal{U}_{\parallel}$, $\mathcal{U}_{\perp}$–$\mathcal{U}_{\dot{s}}$, $\mathcal{U}_{\parallel}$–$\mathcal{U}_{\dot{s}}$); the largest pair also contains $\mathcal{U}_{\parallel}$, so we treat that ordering as descriptive.

5.3 Construct specificity by intervention

Residual coupling alone does not show that each score measures what its name claims. We therefore perturb the reference distribution while keeping the curve and probe windows fixed (Appendix F). The three planned manipulations are late-life holdout$\rightarrow \mathcal{U}_{\parallel}$, cross-subset failure-mode shift$\rightarrow \mathcal{U}_{\perp}$, and operating-condition holdout$\rightarrow \mathcal{U}_{\dot{s}}$. Table 2 reports the outcome: one construct-supported interpretation, one calibration-specificity check, and one unsupported interpretation.

Table 2: **Construct-validity interventions under the final rarity calibration.** Each entry counts the cells in which the intended component has the largest control-normalised response, out of 36 (12 directed subset pairs $\times$ 3 seeds). The curve is fitted once on the reference population and the probe windows are fixed; only the reference distribution used to calibrate U_c changes. The control is a size-matched random block-unit holdout. The full 3×3 confusion matrix, showing which component wins when the intended one does not, is Table 10.

Intervention	Intended component	Selected / 36
Cross-subset failure-mode shift	$\mathcal{U}_{\perp}$	$\mathbf{35/36}$ —*construct-supported*
Late-life holdout	$\mathcal{U}_{\parallel}$	$\mathbf{36/36}$ —*calibration check, near-definitional*
Operating-condition holdout	$\mathcal{U}_{\dot{s}}$	$21/36$ —*not construct-validated*

The three outcomes differ in kind. The key construct result is the cross-subset failure-mode shift, and the response stays asymmetric by probe subset, consistent with failure-mode containment. Removing a failure mode does perturb the transverse marginal, so the row is not self-evidently a construct result: two support artifacts could manufacture it, and the design argues against both. Removing a failure-mode subset perturbs the support of s by $2.6\times$ the control, but the late-life holdout perturbs it nearly three times further still and selects $\mathcal{U}_{\perp}$ in $0/36$, so an s-support artifact has the wrong signature. The same ablation perturbs d-support by $3.8\times$ while producing a transverse response $21\times$ larger, and fitting the dose–response line on the remaining 108 cells leaves all 36 cross-subset cells above it, in the same direction on all 12 directed subset pairs: the response is steeper than its own dose. Those cells are 12 pairs by three seeds rather than 36 independent trials, so we quote counts and effect sizes and not a p-value. Appendix F.4 gives both routes and their limits. Scrambling the transverse residual across units within life-stage bins cuts the hit count by roughly eleven cell-equivalents, so the response is partly carried by unit-level transverse structure; a within-unit scramble moves one cell. We claim no rate interpretation, though the coordinate is not inert (Appendix G). Late-life holdout is $36/36$, but $\mathcal{U}_{\parallel}$ is a tail probability in s and the ablation removes exactly the range of s the probe occupies, so that row is a calibration check whose content is the specificity—the other two did not move—not the diagonal.

6 Held-Out Evaluation Sets

On DS04 and the cross-subset split the conditional statistic is 0.109 ± 0.046 versus 0.096 ± 0.054 on development, reported only as a magnitude because no unit-preserving reference was recomputed there; calibration sensitivity is in Appendix J.

7 Scope, Transfer, and Refusal Across Domains

The unchanged pipeline also runs on four external degradation datasets, with the admissibility gate using the same training-only post-hoc principal curve that defines the downstream PCNG coordinates. Table 3 gives the per-seed outcome: external admissibility is mixed rather than a blanket transfer result, and no external domain earns a GO on any seed.

Table 3: **PCNG admissibility on four external domains under the canonical post-hoc geometry.** CV(L) per seed with its thresholded label (GO < 0.15, MARGINAL $[0.15, 0.35]$, NO-GO > 0.35); CV(L) is the measurement and the label a threshold summary of it. Residual-coupling statistics and unit-preserving references for these domains are in Table 16, and coordinate sensitivity in Table 17.

Domain	CV(L), seeds 42–44			Verdict
PRONOSTIA	0.895	1.459	0.528	NO-GO $\times 3$
XJTU-SY	0.441	0.481	0.759	NO-GO $\times 3$
NASA Battery	0.300	0.329	0.317	MARGINAL $\times 3$
NASA Capacitor	0.377	0.190	0.368	NO-GO/MARGINAL/NO-GO

For PRONOSTIA and XJTU-SY we retain the unit-aware correlation calculations only as diagnostics of reference scaling, not as licensed semantic attribution after a NO-GO gate. Their unit-preserving reference means are 0.254 and 0.167 against pooled observed 0.298 and 0.192, and none of the six runs exceeds its own $p_{97.5}$. That $0/6$ is largely an absence of resolution rather than an absence of coupling: calibrated to these fleets' own geometry and autocorrelation, the diagnostic detects a true coupling of 0.20 in only 10% and 36% of synthetic fleets respectively (Table 9). Most visibly, PRONOSTIA's apparent $7.1\times$ residual-coupling ratio under independent windows becomes $1.17\times$ under a five-test-unit reference (Appendix E).

Outside prognostics, an amended-preregistration CIFAR-10 corruption-transfer check gives a directional construct result (s tracks severity; family shifts select $\mathcal{U}_\perp$ on $2/3$ seeds), not a vision-OOD claim (Appendix M.4).

Capacitor exposes threshold sensitivity, not a malformed fleet. The Capacitor split contains eight training and three test capacitors, with enough repeated windows to estimate per-unit path length and $\dot{s}$. Its seed-level CVs straddle the NO-GO threshold (Table 3), and the corresponding conditional residual-coupling maxima are $0.165, 0.353, 0.231$ (0.249 ± 0.096). The conditional maximum exceeds the raw maximum on only one of three seeds, so we do not use post-gate residual-coupling behaviour as evidence of geometric incoherence, and we report seed-level outcomes rather than a single categorical dataset verdict. Unit bootstrap on the development jobs reproduces a run's own ternary label in only 0.149 of resamples (Appendix H), which is why CV(L) and not the label is treated as the measurement.

What a refusal does and does not mean. A controlled synthetic sweep (Appendix M) measures which violations the admissibility gate is sensitive to. The gate is specifically a test of *path homogeneity*: under non-monotone progression it withdraws while construct accuracy is still 1.00, but curvature and transverse noise never drive $P(\mathrm{GO})$ below 0.5 even as accuracy falls to 0.78 and 0.82 respectively, and under branching the components degrade before the gate objects. So a NO-GO reports elevated path-length heterogeneity rather than a general verdict on geometric reliability, and a GO is correspondingly not a guarantee that the component interpretations are intact. Nor is the gate uniformly restrictive: across the eight encoder groups of Appendix M.3 GO rates span 0.33–1.00 on identical data, five groups earn a GO majority, and our own MonoCNN is among the less admissible (mean CV(L) $= 0.163$, GO rate 0.44, MARGINAL majority), so the gate neither refuses indiscriminately nor privileges the contributed encoder. The pre-registered CIFAR-10-C check refuses on every seed (Appendix M.5).

8 Operational Consequences: Onset and Near-Failure Alarm

We now ask what the licensed transverse component is worth on onset detection and near-failure alarm. Both readouts are label-free at deployment: thresholds come from training pre-onset data; test labels score alarms only.

8.1 Onset detection, and a degeneracy in how it is scored

Onset is where a scalar interface should matter most, and where a ranking metric misleads: SoftCLT-TS2Vec and TS2Vec at d=320 reach area under the receiver-operating-characteristic curve (AUROC)

Table 4: **Near-failure alarm: transverse rarity against classical latent-space novelty scores.** Same frozen embeddings, detector (W=20) and training-pre-onset p_{98} threshold; 12 matched CNN jobs. FA denotes false-alarm rate. *lead$_{first}$* is cycles from first alarm to end of life, so smaller is later and tighter; *first-in-50* is the fraction inside the final 50 cycles. Paired tests: Appendix P.

Score	Coverage	FA ↓	lead$_{\text{first}}$ ↓	first-in-50 ↑
$\mathcal{U}_\perp$ (PCNG), rarity of $R_\perp = d^2/V_\parallel$	0.979 ± 0.072	$\mathbf{0.049 \pm 0.090}$	114.3 ± 71.3	0.312 ± 0.396
$\mathcal{U}_{\text{geo}}$ (secondary), $\bar{d}(1+\alpha\bar{\rho})$	0.979 ± 0.072	0.083 ± 0.107	116.5 ± 88.2	$\mathbf{0.361 \pm 0.396}$
Mahalanobis to train latent Gaussian	0.979 ± 0.072	0.090 ± 0.217	163.6 ± 105.9	0.181 ± 0.313
kNN distance to train z (k=10)	0.847 ± 0.253	0.118 ± 0.168	$\mathbf{102.7 \pm 85.2}$	0.306 ± 0.267
KDE negative log-likelihood (NLL), Scott bandwidth	0.979 ± 0.072	0.167 ± 0.249	184.1 ± 99.9	0.090 ± 0.140
$\lvert h \rvert$ control (Stage-1 health coordinate)	0.979 ± 0.072	0.292 ± 0.309	242.9 ± 94.4	0.062 ± 0.155

values of 0.975–0.995 under CUSUM while their post/pre spike-rate ratios are 1.06 and 1.23 and AUROC falls to 0.58–0.82 under a rolling-rate detector. Reporting a rate statistic or a second detector alongside AUROC separates the two; Appendix N gives the full sweep, in which detector swaps move MonoCNN by 0.027 against 0.322 for the degenerate rows.

8.2 Near-Failure Alarm (CNN)

The near-failure task asks whether the transverse component can identify the near-failure phase without producing excessive alarms during healthy operation. We threshold $U_\perp$ with the same rolling-max detector and training-only threshold on 12 CNN jobs (DS01–DS04, three seeds); $\mathcal{U}_{\text{geo}}$ is a secondary refinement, not a PCNG component. Because $U_\perp$ is ultimately built from distance to a learned curve, a natural challenge is whether the curve adds anything over ordinary latent-space novelty. Mahalanobis distance, kNN distance and kernel-density-estimation (KDE) scores therefore evaluate the same frozen embeddings under identical machinery. A separate question— whether transverse rarity also identifies windows on which the frozen predictor is less accurate—is tested by progression-conditioned selective prediction in Appendix U; that comparison is not supported against random selection.

Coverage is near-ceiling, so false alarms carry the comparison. $U_\perp$ has pooled false-alarm rate 0.049 versus 0.292 for the health control and favours it on all four dataset aggregates, and has a lower false-alarm rate than KDE on $3/4$. Mahalanobis is floor-limited ($11/12$ job ties, nine of them at zero false alarms for both), while kNN alarms later at lower near-failure coverage (0.847 versus 0.979). With four datasets these comparisons are descriptive. We therefore claim improvement over density and health controls, and competitiveness with—not superiority to—distance-based novelty (Appendix P). A matched-Mahalanobis attribution check is PARTIAL (some but not all locked criteria pass) and supports no claim (Appendix V).

9 Conclusion

PCNG turns post-hoc novelty from scalar detection into *falsifiable attribution*: which geometric mode can legitimately be said to make a frozen sequential representation unusual. The transverse coordinate is licensed by intervention, weakened by the scramble designed to destroy its unit-level structure, stable to reference resampling, and useful for a lower-false-alarm near-failure readout. The rate coordinate remains computable but is not construct-validated, unit-preserving analysis licenses no independence claim, and the gate can qualify or refuse a domain before attribution. Negative and unresolved outcomes are outputs of the contract, not exceptions to it.

Scope and limitations. PCNG requires a representation with a defensible dominant progression geometry and does not manufacture one. The gate is informative for path-length heterogeneity but blind to some curvature and transverse-noise failures; it is sensitive both to its thresholds and to the progression coordinate supplying path length; external admissibility is mixed, with no external domain earning a GO on any seed; and the strongest real-data construct validation remains in degradation prognostics. The pre-registered CIFAR-10-C check refuses on every seed (Appendix M.5): a blind gate demonstration, not broad vision validation. Within those limits the contribution is the measurement contract itself, not another novelty score.

AI USE STATEMENT

We used generative AI tools to assist with research ideation and hypothesis refinement, methodological and experimental design discussions, implementation and debugging of experimental code, interpretation of results, and manuscript and LaTeX editing. All AI-assisted contributions were reviewed and verified before inclusion in the final submission.

ETHICS STATEMENT

This work studies degradation prognostics on public engineering benchmarks (N-CMAPSS, PRONOSTIA/FEMTO, XJTU-SY, and the NASA Battery and Capacitor sets). No human subjects, personal data, or personally identifiable information are involved, and no new data were collected. The application domain—predicting remaining useful life of machinery—carries the usual safety consideration that a maintenance decision should not rest on a single scalar score: Section 8.1 shows explicitly that a detector can attain near-perfect ranking metrics while discriminating nothing, and we report rate-based diagnostics alongside AUROC for that reason. We see no other ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

The supplementary material accompanying this submission contains the PCNG implementation, the admissibility gate, the pre-registration files, the stored analysis artefacts, and the scripts that regenerate every table in this paper; a public repository will be released on acceptance. The data splits and evaluation protocol are given in Appendix D, the principal-curve algorithm in Appendix B, and all hyperparameters in Appendix C. All datasets used are public. Embedding packs resolve unit identity from the dedicated `unit_*` fields. Repeated evaluation of the released analysis scripts was deterministic to numerical precision.

REFERENCES

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

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Jose R. Celaya, Chetan S. Kulkarni, Gautam Biswas, Sankalita Saha, and Kai Goebel. A model-based prognostics methodology for electrolytic capacitors based on electrical overstress accelerated aging. In *Annual Conference of the Prognostics and Health Management Society*, 2011. doi: 10.36001/phmconf.2011.v3i1.2004.

Manuel Arias Chao, Chetan Kulkarni, Kai Goebel, and Olga Fink. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, 6(1):5, 2021. doi: 10.3390/data6010005.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607. PMLR, 2020.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting.

In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2352–2359, 2021. doi: 10.24963/ijcai.2021/324.

Olga Fink, Qin Wang, Markus Svensén, Pierre Dersin, Wan-Jui Lee, and Melanie Ducoffe. Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, 92:103678, 2020. doi: 10.1016/j.engappai.2020.103678.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

En Fu, Yanyan Hu, Kaixiang Peng, and Yuxin Chu. Supervised contrastive learning based dual-mixer model for remaining useful life prediction. *Reliability Engineering & System Safety*, 251:110398, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pp. 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1321–1330. PMLR, 2017.

Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989. doi: 10.1080/01621459.1989.10478797.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

Sepp Hochreiter and J"urgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Zhixin Huang, Yujiang He, Chandana Priya Nivarthi, Bernhard Sick, and Christian Gruhl. Time-series representation learning via heterogeneous spatial-temporal contrasting for remaining useful life prediction. In *Pattern Recognition (ICPR 2024)*, volume 15327 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-3-031-78398-2_1.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022. doi: 10.1145/3531146.3533233.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. doi: 10.1145/3442188.3445901.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2nd edition, 2002.

Balázs Kégl, Adam Krzyżak, Tamás Linder, and Kenneth Zeger. Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):281–297, 2000. doi: 10.1109/34.841759.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. doi: 10.1002/aic.690370209.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7167–7177, 2018.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *International Conference on Learning Representations*, 2024.

Yaguo Lei, Naipeng Li, Liang Guo, Ningbo Li, Tao Yan, and Jing Lin. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, 104:799–834, 2018. doi: 10.1016/j.ymssp.2017.11.016.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 (NIPS)*, 2004.

Xiang Li, Qian Ding, and Jian-Qiao Sun. Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172:1–11, 2018. doi: 10.1016/j.ress.2017.11.021.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? A meta review of evaluation failures across machine learning. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Don H. Loftsgaarden and Charles P. Quesenberry. A nonparametric estimate of a multivariate probability density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965. doi: 10.1214/aoms/1177700077.

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 7045–7054. PMLR, 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019.

Patrick Nectoux, Rafael Gouriveau, Kamal Medjaher, Emmanuel Ramasso, Brigitte Morello, Noureddine Zerhouni, and Christophe Varnier. Pronostia: An experimental platform for bearings accelerated degradation tests. In *IEEE International Conference on Prognostics and Health Management*, pp. 1–8, 2012.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Umut Ozertem and Deniz Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286, 2011.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Emmanuel Ramasso and Abhinav Saxena. Performance benchmarking and analysis of prognostic methods for CMAPSS datasets. *International Journal of Prognostics and Health Management*, 5 (2), 2014. doi: 10.36001/ijphm.2014.v5i2.2236.

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 32 (NeurIPS)*, 2019.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007. doi: 10.5281/zenodo.40328.

Bhaskar Saha and Kai Goebel. Battery data set. NASA Prognostics Data Repository, 2007.

Abhinav Saxena and Kai Goebel. Turbofan engine degradation simulation data set. NASA Prognostics Data Repository, 2008.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, pp. 20827–20840, 2022.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. I-glide: Input groups for latent health indicators in degradation estimation. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2025)*, pp. 395–411, 2025. doi: 10.1007/978-3-032-06106-5_23.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Step: Learning structured embeddings for progressive time series. *arXiv preprint arXiv:2605.31061*, 2026.

Robert Tibshirani. Principal curves revisited. *Statistics and Computing*, 2(4):183–190, 1992. doi: 10.1007/BF01889678.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jakob J. Verbeek, Nikos Vlassis, and Ben Kröse. A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8):1009–1017, 2002. doi: 10.1016/S0167-8655(02)00032-6.

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

Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 69(1):401–412, 2020. doi: 10.1109/TR.2018.2882682.

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)*, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. doi: 10.1007/s11263-024-02117-4.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8980–8987, 2022. doi: 10.1609/aaai.v36i8.20881.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 3988–4003, 2022.

Mengjie Zhao and Olga Fink. Disentangling slow and fast temporal dynamics in degradation inference with hierarchical differential models. *arXiv preprint arXiv:2509.00639*, 2025.

Shuai Zheng, Kosta Ristovski, Ahmed Farahat, and Chetan Gupta. Long short-term memory network for remaining useful life estimation. In *IEEE International Conference on Prognostics and Health Management (ICPHM)*, pp. 88–95, 2017. doi: 10.1109/ICPHM.2017.7998311.

Appendix roadmap. The appendices are grouped by the claim they support, and within a group by the order the main text raises them. Appendix A extends Section 2 to the neighbouring literatures—principal curves, novelty detection, uncertainty quantification, selective prediction, measurement validity, and time-series representation learning for prognostics—stating for each what PCNG borrows and where it departs. Appendices B–D give the method and everything needed to reproduce it. Appendices E–L support the correlation analysis of Sections 5 and 6, including the rate-intervention and residual-coupling tests that come back negative and the synthetic positive control of Appendix G. Appendix M measures the operating characteristic of the PCNG admissibility gate, including where it is blind, and includes the controlled CIFAR-10 corruption-transfer and pre-registered CIFAR-10-C refusal checks. Appendices N–U support the detection results of Section 8, ending with the progression-conditioned selective-prediction test, whose result is a null. Appendix V reports a PARTIAL matched-Mahalanobis attribution check. Appendix W gives a schematic of backbone-dependent transverse regimes.

A Extended Related Work

Section 2 is constrained by the page limit to the work PCNG is most directly built on. This appendix places the protocol in its wider setting and states, for each neighbouring literature, what PCNG borrows and where it departs.

A.1 Principal curves and the geometry of learned representations

The principal curve is a self-consistent one-dimensional summary of a distribution (Hastie & Stuetzle, 1989): a curve that passes through the middle of the data in the sense that each point on it is the conditional mean of the points projecting there. The definition is elegant but the estimator is not unique, and the subsequent literature is largely a sequence of ways to make it well-posed. Tibshirani (1992) recast the curve as the solution of a mixture model, removing the bias of the original conditional-mean construction. Kégl et al. (2000) replaced the smoothness prior with a length constraint and gave learning-theoretic guarantees together with a polygonal-line algorithm; Verbeek et al. (2002) built the curve incrementally from local line segments, which behaves better on data with loops or branches. Ozertem & Erdogmus (2011) abandoned self-consistency for a differential-geometric definition—the curve as a ridge of the density—which makes the object local and removes the dependence on a global regularisation parameter.

We fit the curve with the conditional-mean iteration of Hastie & Stuetzle (1989) initialised by the first principal component (Jolliffe, 2002), and add a soft temporal-order term (Appendix B) because our data carries a monotone covariate the unsupervised estimators cannot use. What we take from this literature is not an algorithm but a stance—the one-dimensional summary and the residual off it are different objects and deserve separate readouts.

The wider question of what geometry a learned representation has is usually asked through dimension rather than through curves. Classical manifold learning recovers a global low-dimensional embedding (Tenenbaum et al., 2000), and nonlinear principal component analysis (PCA) via autoassociative networks (Kramer, 1991) is the direct ancestor of the autoencoder baselines in Section 4. Intrinsic dimension estimated by maximum likelihood (Levina & Bickel, 2004) has since been used to show that representations compress to a low-dimensional manifold as depth increases (Ansuini et al., 2019), and that the intrinsic dimension of image data predicts how hard it is to learn from (Pope et al., 2021). Moor et al. (2020) go further and regularise the representation so that its topology matches the input's. All of these summarise geometry with a *scalar* describing the population. PCNG asks a different question: not how many dimensions the population occupies, but where one window sits relative to the population along interpretable axes. Earlier revisions of this paper reported an effective rank (Roy & Vetterli, 2007) of the residualised system; we have withdrawn it, since it is a deterministic function of the same three correlations and therefore inherits whatever they establish, which here is nothing.

A.2 Novelty and out-of-distribution detection

The modern OOD literature begins with the observation that a softmax score separates in- from out-of-distribution inputs better than chance (Hendrycks & Gimpel, 2017), and proceeds by sharpening that signal—temperature scaling and input perturbation (Liang et al., 2018), class-conditional Gaussian

distances in feature space (Lee et al., 2018), non-parametric kNN distances to the training set (Sun et al., 2022), and likelihood ratios that cancel the background component making raw generative likelihoods unreliable (Ren et al., 2019). That last problem is worth restating, because it is the closest thing in this literature to the failure PCNG is designed around: deep generative models can assign *higher* likelihood to out-of-distribution inputs than to their own training data (Nalisnick et al., 2019), so a single density-derived scalar can be confidently wrong in a way no threshold repairs. Yang et al. (2024) survey the field and, usefully for us, separate the sub-problems—covariate shift, semantic novelty, anomaly—that one detector silently conflates.

Three of these scores are baselines in Section 8.2, run on the identical frozen embeddings under the identical detector, and the summary of that comparison is in Appendix P: the curve beats density-based novelty and the health coordinate; the Mahalanobis comparison is floor-limited. What PCNG offers there is not a lower number but a decomposed one—a distance-to-manifold, a position along it and a rate, each separately calibrated and separately interpretable—plus a gate that can decline the fit. We do not claim the decomposition is worth more than the tie; we claim it answers questions the comparison is uninformative.

Detecting shift at the level of a *dataset* rather than a point is a parallel line (Rabanser et al., 2019), and its finding that dimensionality-reduced two-sample tests outperform more elaborate alternatives is consistent with our own experience that simple geometry travels further than it should.

A.3 Uncertainty quantification, and why PCNG is not that

The uncertainty literature supplies the scalar PCNG is reacting to. Monte-Carlo dropout (Gal & Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017) and heteroscedastic regression (Kendall & Gal, 2017) each yield a number intended to say how much to trust a prediction, and the aleatoric/epistemic distinction gives that number a taxonomy (Hüllermeier & Waegeman, 2021). Two results bound what the taxonomy delivers in practice: modern networks are badly calibrated (Guo et al., 2017), and methods that look best in distribution degrade unevenly under shift (Ovadia et al., 2019). Conformal prediction sidesteps calibration with distribution-free finite-sample guarantees (Vovk et al., 2005; Angelopoulos & Bates, 2023), at the cost of returning an interval rather than an explanation.

PCNG is deliberately not in this family, and the Introduction says so. It produces no predictive interval, no coverage guarantee and no variance estimate, and Appendix U reports the experiment that tests the natural bridge between the two—whether transverse rarity flags windows where the frozen predictor is less accurate—and finds a null against a random baseline. The relationship we do claim is upstream of uncertainty: an epistemic-uncertainty estimate is supposed to be large when the input is unlike the training population, and PCNG measures *in what way* it is unlike it, which is a property of the representation rather than of the predictor.

A.4 Selective prediction and the evaluation of it

The risk–coverage curve and the formal treatment of a predictor with an abstention option are due to El-Yaniv & Wiener (2010), extended to deep networks by Geifman & El-Yaniv (2017). The summary statistics used in Appendix U—the area under the risk–coverage curve and its excess over an oracle ranking—are introduced by Geifman et al. (2019). Ding et al. (2020) and, more recently, Traub et al. (2024) document how easily this evaluation misleads: a selective-prediction comparison inherits every confound of the underlying task, and an impressive sweep against a poorly chosen reference says nothing. That is precisely the trap our own experiment falls into if read carelessly, which is why Appendix U reports the $12/0/0$ sweep against progression *and* the fact that progression is worse than random selection once life stage is controlled.

A.5 Measurement and construct validity

PCNG is presented as a measurement protocol, and that word is doing real work. The argument that machine-learning quantities should be treated as measurements of constructs—with the reliability and validity obligations that follow—is made by Jacobs & Wallach (2021), and the gap between what benchmarks are taken to measure and what they do measure is documented across the field (Raji et al., 2021; Liao et al., 2021; Hutchinson et al., 2022). The practical consequence here is the

structure of Section 5 and Table 1: we do not assert that $\mathcal{U}_\perp$ measures transverse deviation because we computed it from a transverse distance. We intervene on the geometry and check that the intended component responds (Appendix F), attempt to falsify the result by scrambling the residual across units (Appendix I), resample the reference to see whether the decision is stable (Appendix H), and report the rate intervention and the residual-coupling diagnostic—the two tests that come back negative—alongside those that do not.

A.6 Self-supervised representation learning for time series

The literature baselines PCNG reads are drawn from this literature and are used unmodified. Contrastive predictive coding established the template of learning representations by predicting future latents (van den Oord et al., 2018); triplet losses with time-based negative sampling adapted it to multivariate series (Franceschi et al., 2019); temporal neighbourhood coding (Tonekaboni et al., 2021) and temporal–contextual contrasting (Eldele et al., 2021) defined positives through temporal proximity and augmentation consistency respectively; time–frequency consistency (Zhang et al., 2022) added a spectral view for transfer across domains. The encoders we freeze and read—SimCLR-style augmentation contrast (Chen et al., 2020), hierarchical contrastive learning over augmented contexts (Yue et al., 2022), and its soft-assignment variant (Lee et al., 2024)—are drawn from this family and used unmodified.

Two relevant architecture families we did not include are patch-based transformers such as PatchTST (Nie et al., 2023) and general-purpose temporal backbones such as TimesNet (Wu et al., 2023). Both support representation learning as well as forecasting or classification, but they fall outside the frozen-encoder benchmark assembled here. Evaluating them under the same PCNG protocol would extend backbone coverage; their omission does not enter the claims tested in this paper.

A.7 Prognostics and the datasets used here

Data-driven prognostics has a well-surveyed pipeline from acquisition to RUL (Lei et al., 2018), and the deep-learning turn within it is reviewed with its open problems by Fink et al. (2020). On C-MAPSS the reference architectures are a deep convolutional regressor (Li et al., 2018) and a long short-term memory (LSTM) regressor (Zheng et al., 2017), and Ramasso & Saxena (2014) provide the benchmarking analysis that makes cross-paper comparison on those datasets interpretable at all—including the observation that reported results are often not comparable because the evaluation protocol differs. N-CMAPSS (Chao et al., 2021) is the successor benchmark we use for development, with PRONOSTIA/FEMTO (Nectoux et al., 2012), XJTU-SY (Wang et al., 2020) and the NASA battery (Saha & Goebel, 2007) and capacitor (Celaya et al., 2011) sets as external domains.

Contrastive and self-supervised representation learning has begun to reach this domain directly—supervised contrastive objectives for RUL regression (Fu et al., 2024), and heterogeneous spatio-temporal contrasting for the same task (Huang et al., 2025). Both aim at prediction accuracy. PCNG does not compete with them: it is a post-hoc readout that takes any such frozen encoder as input, and Section 5 deliberately spans thirteen of them rather than proposing a fourteenth.

B Principal Curve Algorithm

The curve is initialized with the first PCA component of the training embeddings (Jolliffe, 2002), then refined by 20 alternating iterations of (1) a soft temporal-order projection (β=0.1) and (2) a kernel-smoothed conditional-mean update. The stopping criterion is $\|\gamma^{(k+1)} - \gamma^{(k)}\|_F < 10^{-4}$. The local kNN density term with k=10 and α=1.0 is used only by the operational $\mathcal{U}_{\text{geo}}$ statistic; PCNG component rarity uses empirical reference CDFs. All calibration statistics are computed on the training reference only.

B.1 Rarity calibration

This expands the compressed statement in Section 3.2. For each component c, let $\hat{F}_R^{(c)}$ be the empirical CDF of its raw score on the training reference. The tail probabilities entering Equation 7

are

$$q_\perp(r) = 1 - \hat{F}_R^{(\perp)}(r), \qquad q_c(r) = 2\min\{\hat{F}_R^{(c)}(r), 1 - \hat{F}_R^{(c)}(r)\}, \quad c \in \{\|, \dot{s}\}. \tag{9}$$

The $1/N_R$ floor limits the reported rarity at the resolution of the reference sample and avoids arbitrary extrapolation beyond empirical support.

The transverse scale $V_\| = \mathrm{Var}(s_{\mathrm{train}})$ is a sample variance with ddof$=1$, held fixed across conditions in the construct ablations. Because the rarity map is a tail probability of a monotone transform of d, every reported quantity is invariant to the value of $V_\|$: it sets the units of $R_\perp$ but not the rank of any window, and $U_\perp$ depends on $R_\perp$ only through its position in the reference distribution.

C Implementation Details

Tables 5 and 6 list every setting needed to reproduce the reported results. All values are held fixed across datasets, backbones, and latent dimensions unless the table states otherwise; nothing is tuned per subset or per seed. The Stage-1 objective is the fixed weighted sum

$$\mathcal{L}_{\mathrm{rep}} = \sum_{k \in \mathcal{K}} \lambda_k \mathcal{L}_k, \qquad \mathcal{K} = \{\mathrm{ssl},\ \mathrm{temp},\ \mathrm{mono},\ \mathrm{pr},\ \mathrm{spread}\}, \tag{10}$$

with the weights below.

Table 5: **Implementation configuration, Stage 1.** The weights λ_k enter Equation (10) and are held fixed across all datasets, backbones, and latent dimensions.

Stage 1 — encoder and optimization	
CNN×2 encoder (Ours / MonoCNN)	Residual depthwise–pointwise CNN, width×2 (channels $16/32/64$), 4 temporal tokens with dual GAP+token pooling, multi-scale late block, SE off; linear latent $z \in \mathbb{R}^d$; ≈9.1K parameters (N-CMAPSS T=1024, C=14)
TR×1 encoder (Ours / MonoTR)	single Transformer encoder block, same projection to $\mathbb{R}^d$; used for mechanism validation only
Health head g_ϕ	Linear($d \to 1$) with sigmoid output
Projection head q_ψ	used in Stage 1 only; discarded before Stage 2
Window length T, batch size, epochs	1024, 256, 50
Optimizer, learning rate	AdamW, base 10^{-3} with cosine decay (warmup 2 epochs; floor 10^{-5}); weight decay 10^{-4}
Temperature τ (InfoNCE / TEMP)	0.10
Augmentations	additive noise + light jitter (two independent views of the same window; no temporal mask / roll by default)
Latent dimensions d	$\{2, 4, 8, 16, 32\}$
Seeds	42, 43, 44
Stage 1 — loss weights (Equation (10))	
λ_{ssl} (contrastive)	0.30
λ_{temp} (temporal consistency)	0.10
λ_{mono} (monotone health ordering)	0.80 *(largest)*
λ_{pr} (pairwise representation ordering)	0.40
$\lambda_{\mathrm{spread}}$ (health spread floor)	0.30
Mono / PR margins; spread floor	0.05, 0.05, 0.15
Stage 1 — pair construction	
Temporal positives $\mathcal{P}_T$	$0 < \lvert\Delta\rvert \le k_{\mathrm{pos}}$ with k_{pos}=2; Gaussian weight σ_t=1.0; representation-similarity gate 0.5
Ranking / PR pairs $\mathcal{P}_R$	minimum separation $\Delta t_{\min}$=4; PR far offsets in $[32, 48]$ (batch-size auto-hparams on N-CMAPSS)
Negative set $\mathcal{A}(i)$	same-unit temporal near-neighbours excluded ($\lvert\Delta\rvert \le 2$)
All pair sets	built from same-unit windows of the training split only

Table 6: **Implementation configuration, Stages 2–3 and evaluation.** Every quantity is fitted or selected on training data only; no test trajectory influences any of them.

Stage 2 — post-hoc geometry	
Principal curve γ	initialized at PCA-1 of the training embeddings; 20 alternating iterations
Order regularizer, stopping rule	β=0.1; $\|\gamma^{(k+1)} - \gamma^{(k)}\|_F < 10^{-4}$
Local density term ($\mathcal{U}_{\text{geo}}$ ablation)	kNN log-density, k=10
PCNG rarity calibration	empirical CDF; upper tail for $U_\perp$, two-sided tails for $U_\parallel$ and $U_{\dot{s}}$
Novelty score $\mathcal{U}_{\text{geo}}$	α=1.0; d and ρ normalized by their training-set 99th percentiles
Stage 3 — task readouts	
Onset detector	$X_t = \mathbf{1}[h_t > \tau_{95}]$ with τ_{95} from the training percentile, then Bernoulli CUSUM
Onset baseline score	Mahalanobis distance with (μ, Σ) from the pre-onset training representations, same CUSUM
Near-failure alarm	rolling maximum of $U_\perp$ over W=20 windows, training-set pre-onset p_{98} threshold; $\mathcal{U}_{\text{geo}}$ reported as a secondary ablation
Evaluation protocol	
Splits	unit-held-out; predefined N-CMAPSS partitions; no validation set
Fitted on training data only	encoder, principal curve, PCNG reference CDFs, percentile normalization, PCA sign, linear heads, alarm thresholds
Reported statistics	mean±std over seeds; median per-unit Spearman for ρ

D Data Splits and Evaluation Protocol

The primary experiments use N-CMAPSS (Chao et al., 2021). Each trajectory is treated as a unit, and all windows from a unit are kept within the same split. The benchmark uses the predefined unit-level training and test partitions for each experiment, and no validation set is used, so the held-out unit identities follow from the benchmark rather than from a choice of ours and are reproducible without a list. Their *counts* are load-bearing for the reference of Appendix E, whose width is set by the number of held-out units rather than the number of windows, and we state them: four test units on DS01, three on DS02, six on DS03, five bearings on PRONOSTIA and three on XJTU-SY; NASA Capacitor has eight training and three test capacitors, and NASA Battery three training cells and one test cell. The test split is never used for checkpoint selection, geometry fitting, normalization, threshold calibration, or model selection. The cross-subset experiment is reported separately and uses the training subset specified by the corresponding evaluation protocol.

E Size-Matched Nulls for the Conditional Rank-Residual Statistics

The residual-coupling statistic reported in Section 5 has a reference distribution that depends on n_{test}: residualising a rank-transformed variable within quantile bins of s consumes degrees of freedom, and the residual correlation between genuinely independent variables shrinks as n grows. A value of $\max|\rho| = 0.10$ is decisive evidence of dependence at $n = 4000$ and entirely unremarkable at $n = 100$. We therefore never compare an observed value to a fixed threshold.

Table 7 gives the null distribution, simulated by drawing three independent standard normal variables and an independent uniform s at each sample size and passing them through the identical residualisation and binning used on real data (400 replicates per row, 12 bins).

What this table controls for, and what it assumes. The reference above controls correctly for sample size and for the degrees of freedom the residualisation consumes—that is a real effect and ignoring it would be worse—but it assumes exchangeable windows. Degradation windows are nested within a small number of trajectories: five engines of three hundred windows each are not fifteen hundred independent observations, because window 101 resembles windows 100 and 102 and because one engine may sit systematically above another throughout its life. A reference drawn as

Table 7: **Null distribution of the conditional rank-residual correlation** under independence, by sample size. "single pair" is the statistic for one designated pair; "max over 3" is the maximum over the three pairs, which is what the conditional rank-residual $\max|\rho|$ reports and is accordingly larger. $p_{97.5}$ is the one-sided 97.5th percentile.

	single pair		max over 3 pairs	
n_{test}	mean	$p_{97.5}$	mean	$p_{97.5}$
50	0.129	0.328	0.215	0.424
100	0.092	0.233	0.145	0.274
200	0.062	0.170	0.100	0.187
300	0.048	0.136	0.079	0.150
500	0.035	0.100	0.061	0.119
800	0.028	0.081	0.047	0.097
1200	0.025	0.065	0.039	0.079
1726	0.020	0.055	0.032	0.064
2500	0.016	0.045	0.027	0.052
4000	0.013	0.034	0.021	0.041

independent rows therefore describes a world with far more effective independence than the data have, and a statistic read against it looks larger than it is.

We instead use a unit-preserving randomisation that retains within-unit temporal structure and unit-specific offsets while disrupting cross-coordinate alignment. Every window is still used. For each held-out unit, each coordinate's s-residual series is rotated by an independent random offset, which leaves its marginal and unit level unchanged, preserves its circular autocorrelation structure, and removes the contemporaneous pairing between coordinates. The question the reference answers changes accordingly: not *is this more correlated than independent windows would be*, but *is this more coupled than the dependence already inside real units produces*. Table 8 reports it on the fifteen runs we recomputed.

Table 8: **Unit-preserving residual-coupling reference**, 2000 within-unit circular shifts per run. "unit ref." is the mean null maximum; exceedance is run-specific against that run's own $p_{97.5}$. The unit-aware reference is 1.2–6.5× wider than the matched independent-row reference across the 15 runs.

Runs	test units	n_{test}	observed	indep. ref.	unit ref.	$> p_{97.5}$
MonoCNN DS01–DS03 (9)	3–6	1076–2732	0.056–0.169	0.025–0.041	0.049	4/9
PRONOSTIA (3)	5	1000	0.212–0.425	0.041–0.043	0.254	0/3
XJTU-SY (3)	3	294–558	0.188–0.196	0.055–0.077	0.167	0/3
Pooled (15)	3–6	—	0.056–0.425	—	0.113	**4/15**

The per-run widening relative to the independent-row reference ranges from 1.15× to 6.52×. PRONOSTIA remains the clearest calibration example: the pooled observed mean is 0.298, about 7.10× an independent-row reference of 0.042, but only 1.17× the pooled unit-preserving mean of 0.254. We use this contrast to demonstrate the cost of window-level exchangeability, not as licensed semantic attribution, because the canonical post-hoc admissibility gate declines PRONOSTIA on all three seeds.

How much coupling this diagnostic would have found. We retuned the synthetic resolution check to the unit-preserving reference, separately for each fleet geometry, and applied the same run-level $p_{97.5}$ decision. Table 9 shows that the diagnostic's resolution is a property of the fleet rather than of the statistic. On N-CMAPSS-like fleets it detects an injected coupling of $\rho = 0.20$ in 92% of synthetic fleets and $\rho = 0.10$ in 56%. The ρ=0 simulation gives a per-run false-positive rate of 0.05, confirming calibration at the run level; because the nine empirical runs are not independent replicates, we do not convert the four exceedances into a fleet-level probability or significance claim. On PRONOSTIA-like fleets the same design detects $\rho = 0.20$ only 10% of the time, and on XJTU-like fleets 36%: with 200 or fewer windows per unit and, for PRONOSTIA, a residual lag-1 of 0.94, a

circular shift preserves each series' slow structure, so few effectively distinct alignments exist and the reference is correspondingly wide. The consequence is asymmetric: the in-domain result is a measurement, whereas $0/6$ on the external fleets is mostly an absence of resolution and should not be read as an absence of coupling.

Table 9: **Audited resolution of the unit-preserving residual-coupling diagnostic, by fleet geometry.** Each row is calibrated to one fleet's own unit count, windows per unit, residual lag-1 autocorrelation and reference width; the per-unit offset scale is tuned by bisection until the synthetic circular-shift reference mean matches the observed one. Detection is the fraction of 100 synthetic fleets whose statistic exceeds its own $p_{97.5}$ under 150 shifts. The ρ=0 column is the false-positive rate and is the calibration check. The simulation is a design-resolution check, not a confidence statement about the empirical jobs.

					detection at true coupling ρ		
Fleet geometry	units	win./unit	lag1	unit ref.	ρ=0	ρ=0.10	ρ=0.20
N-CMAPSS-like	4	431	0.13	0.049	0.05	**0.56**	**0.92**
PRONOSTIA-like	5	200	0.94	0.254	0.02	0.04	0.10
XJTU-like	3	150	0.05	0.167	0.02	0.12	0.36

The nine empirical N-CMAPSS jobs remain three seeds on three benchmark subsets, not nine independent fleet replicates, so we do not exponentiate the per-run miss rate into a fleet-level probability or confidence bound. The synthetic model is stylised and is used only to show that moderate residual coupling is resolvable under the unit-preserving reference.

F Construct-Validity Ablation: Protocol and Full Results

Table 2 in Section 5.3 reports three different outcomes under the final $R_c \to U_c$ calibration, and we keep them distinct: the cross-subset manipulation gives construct support for $\mathcal{U}_\perp$; the late-life manipulation gives a $36/36$ calibration-specificity check for $\mathcal{U}_\parallel$ that is close to definitional rather than the discovery of an independent construct; and the rate manipulation remains weak, so $\mathcal{U}_{\dot{s}}$ is not construct-validated. The protocol below keeps the geometric frame fixed and changes only the reference distribution used to calibrate rarity, and reports the full confusion matrix behind those counts.

F.1 Protocol

The principal curve is fitted *once* on the full training population and never refitted. The probe set P—all held-out test windows—is fixed across every condition. The only change between conditions is the reference set R used to estimate the empirical CDFs in Eq. 7. For an ablation A that removes one stratum from the reference, the response of component k is

$$\mathrm{resp}_k(A) = \mathbb{E}_P\big[\mathcal{U}_k\big(P \mid R \setminus A\big) - \mathcal{U}_k\big(P \mid R_{\mathrm{full}}\big)\big]\,, \tag{11}$$

with a bootstrap over probe units. Because the probe is identical on both sides, there is no positional confound in the response difference. Because the curve is fixed, reference ablation cannot manufacture transverse deviation by moving the projection path.

The three planned manipulations are *cross-subset failure-mode shift*$\to U_\perp$, *late-life holdout*$\to U_\parallel$, and *operating-condition holdout*$\to U_{\dot{s}}$. The control is a size-matched random block-unit holdout. Decisions use control-normalised responses $\tilde{r}_k = r_k / \max(|r_k^{\mathrm{ctrl}}|, \varepsilon)$ so that component-specific baseline reactivity does not determine the argmax.

F.2 Rejected design alternatives

Two earlier designs were rejected. Refitting the curve on each ablated reference created an artificial end-of-curve transverse response when late-life windows were removed. Window-level thinning was also a poor control because it perturbs the reference differently from block-unit removal. The final protocol fixes the curve and uses block-unit controls. It also calibrates each raw geometric coordinate to dimensionless empirical rarity, avoiding coordinate-unit dependence.

F.3 WHICH COMPONENT ACTUALLY WINS: THE FULL 3×3

Taking an argmax means all three control-normalised responses exist in every cell, so the off-diagonal is available and we report it. The question it answers is whether $\mathcal{U}_\perp$ is *specifically* responsive to the cross-subset manipulation or simply the most reactive component under any reference ablation.

Table 10: **Construct-validity confusion matrix.** Rows are interventions, columns are the component with the largest control-normalised response; each row is 36 cells (12 directed subset pairs $\times$ 3 seeds). Diagonal entries are the counts reported in Table 2. No cell is a tie.

Intervention	$\mathcal{U}_\perp$ wins	$\mathcal{U}_\parallel$ wins	$\mathcal{U}_{\dot{s}}$ wins
Cross-subset failure-mode shift	**35**	1	0
Late-life holdout	0	**36**	0
Operating-condition holdout	3	12	**21**

Table 11 gives the underlying mean responses. $\mathcal{U}_\perp$ wins 35 cells on its own row, 0 on the late-life row and 3 on the operating-condition row. The transverse result is therefore selective, not generic reactivity. The operating-condition row is also informative about what we decline to claim: when the rate manipulation fails to select $\mathcal{U}_{\dot{s}}$ it selects $\mathcal{U}_\parallel$ twelve times and $\mathcal{U}_\perp$ three, so the failure is a confusion with progression rather than a scatter into the transverse coordinate.

Table 11: **Mean responses per intervention**, in raw nats (left) and after control normalisation (right), averaged over the 36 cells of each row. Raw nats are interpretable units of log-rarity; the normalised values are ratios whose denominators can be small, and are reported because they, not the raw values, determine the argmax.

	raw response (nats)			$\tilde{r}$		
Intervention	$\mathcal{U}_\perp$	$\mathcal{U}_\parallel$	$\mathcal{U}_{\dot{s}}$	$\mathcal{U}_\perp$	$\mathcal{U}_\parallel$	$\mathcal{U}_{\dot{s}}$
Cross-subset	$+\mathbf{1.302}$	$+0.080$	-0.095	$+\mathbf{34.92}$	$+4.23$	-3.03
Late-life	$+0.165$	$+\mathbf{2.565}$	$+0.027$	-2.73	$+\mathbf{129.23}$	-0.01
Operating cond.	-0.084	-0.011	$+\mathbf{0.037}$	-3.66	-0.65	$+\mathbf{1.22}$

Margins, and the one cell that misses. With $\tilde{r}_{(1)} - \tilde{r}_{(2)}$ as the winning margin, the cross-subset row has median 6.44 (interquartile range (IQR) $[3.09, 29.60]$, minimum 0.53) and no cell below 0.05; the late-life row has median 50.27 and none below 0.05; the operating-condition row has median 0.92 and two cells below 0.05, one of them a hit decided by 0.021. The transverse result is not won by hair's-breadth margins and the rate result partly is.

The single cross-subset miss is `ds04`→`ds02`, seed 44. Its raw responses are $+2.092$, $+0.227$ and -0.161 nats, so $\mathcal{U}_\perp$ dominates by roughly nine to one; its control responses are -0.179, -0.0068 and $+0.028$. Because normalisation divides by $\max(10^{-6},\, 0.25 \cdot \mathrm{mean}_j |r_j^{\mathrm{ctrl}}|)$ and the $\mathcal{U}_\parallel$ control falls under that floor, $\tilde{r}$ becomes $+11.72$ against $+12.72$ and the argmax flips. On raw responses the cross-subset row is $36/36$. We report $35/36$ because control normalisation is the pre-specified statistic and changing it after seeing the result is not a choice this paper makes; the raw figure is stated here so the reader can see that the miss is a normalisation edge case rather than a substantive failure. This is also the cell that moves the decision boundary under reference resampling (Appendix H) and the sole D_0 miss in the scramble control (Appendix I): one identifiable weak configuration rather than diffuse instability.

F.4 DOES THE ABLATION MANUFACTURE THE RESPONSE BY SHIFTING SUPPORT?

Removing a failure-mode-defined subset from the reference is not guaranteed to perturb the reference the same way removing a random block of units does, even at matched unit count. If it shifts the support of s or of d more than the control does, that shift—rather than failure-mode structure—could produce the response. We measure both directly in Table 12, as the two-sample Kolmogorov–Smirnov distance between the reference before and after each ablation.

Table 12: **Reference support perturbation per ablation**, mean over 36 cells. KS denotes the two-sample Kolmogorov–Smirnov distance. The control removes a size-matched random block of units.

Ablation	KS(s)	KS(d)	units removed
Cross-subset failure-mode shift	0.151	0.228	6.3
Late-life holdout	0.405	0.133	12.5
Operating-condition holdout	0.048	0.050	3.9
Control (random block-unit)	0.059	0.060	6.6

Both routes are open on this evidence: the cross-subset ablation shifts s by $2.6\times$ and d by $3.8\times$ the control. We address them separately.

The s-support route has the wrong signature. The late-life holdout is a large and deliberate s-support ablation—KS(s) of 0.405, nearly three times the cross-subset value—and under it $\mathcal{U}_\perp$ wins $0/36$ while $\mathcal{U}_\parallel$ wins $36/36$ (Table 10). Perturbing the support of s, even severely, selects the progression coordinate and not the transverse one. The cross-subset ablation perturbs s far less and selects $\mathcal{U}_\perp$ in $35/36$. An s-support artifact does not have that signature.

The d-support route is insufficient to explain the response. $\mathcal{U}_\perp$ is a rarity of transverse distance, so d-support is the route that could plausibly manufacture it, and it is the larger of the two gaps. The descriptive comparison we use needs no additional model: relative to the control the cross-subset ablation perturbs d-support by $3.8\times$ (0.228 against 0.060) and produces a transverse response $21\times$ larger ($+1.302$ against $+0.061$ nats); relative to the late-life ablation it perturbs d-support by $1.7\times$ and responds $7.9\times$. A response that scaled with d-support perturbation would not be five to six times steeper than its own dose in both comparisons.

Fitting a line on the control, operating-condition and late-life cells ($n = 108$) gives $\mathrm{resp}_{\mathcal{U}_\perp} = -0.059 + 1.318\,\mathrm{KS}(d)$, which predicts $+0.242$ nats at the cross-subset dose against $+1.302$ observed. All 36 cross-subset cells lie above that line, with the same sign on all 12 directed pairs; the mean residual is $+1.061$ nats (median $+1.093$, standard deviation (SD) 0.636). We deliberately attach no p-value to that count. A one-sided sign test over 36 cells, or a one-sample t-test on the 36 residuals, would treat 12 directed subset pairs by three seeds as 36 independent draws and return a spuriously tiny figure; that is the same pseudo-replication we decline in Appendix I and in the operational comparison of Section 8.2, and it would be inconsistent to accept it here merely because the number would be favourable. The 12 pairs are not mutually independent either—four subsets generate all of them, so each subset appears in six—so a pair-level test would not repair the problem. The count and the effect size are the claim. We also note that the fit explains only four per cent of the variance over the range it is fitted on, which is itself evidence that d-support perturbation is a weak driver of the transverse response, and that the cross-subset dose lies above that range, so the residual is reported as a second check on the dose-response comparison rather than as the primary argument.

F.5 What this appendix does and does not establish

The $U_\parallel$ response is selected in all $36/36$ late-life cells and the $\mathcal{U}_\perp$ response in $35/36$ cross-subset cells, with wide margins, a specific off-diagonal, and both support routes accounted for. The rate manipulation selects $\mathcal{U}_{\dot{s}}$ in only $21/36$ with two near-ties, so we do not claim a construct-valid rate interpretation. These are argmax-response summaries, not tests of independence between the three components.

Appendix G runs the complementary check on synthetic data with known generative factors. The cross-subset $\mathcal{U}_\perp$ result is the stronger transverse construct test because its probe/reference roles are directional. The late-life result is close to definitional: removing late-life reference windows directly changes the support of s, so its value is specificity to $\mathcal{U}_\parallel$, not evidence that the geometry has discovered time itself.

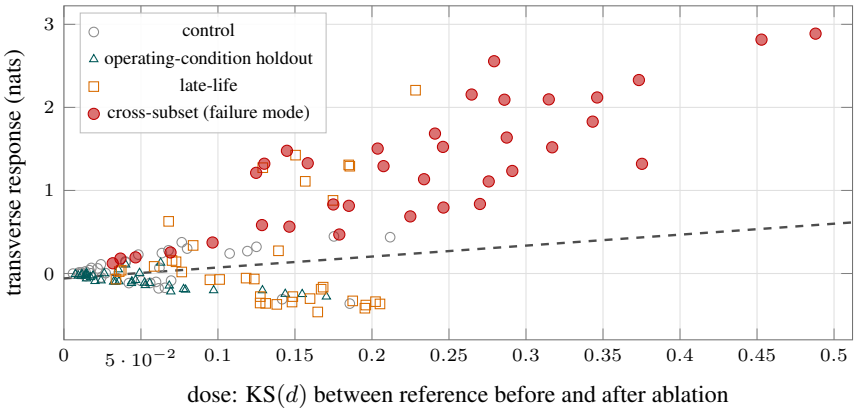

Figure 2: **The transverse response is steeper than its own dose.** Each point is one cell: the horizontal axis is how much the ablation moves the reference distribution of d (a two-sample KS distance), the vertical axis the resulting response of $\mathcal{U}_\perp$. The dashed line is fitted on the control, operating-condition and late-life cells only ($n = 108$, $R^2 = 0.039$). If the cross-subset response followed the linear dose–response relationship observed in the other ablations, the cells would lie near that line; instead all 36 of them lie above it (mean residual $+1.061$ nats, in the same direction on all 12 directed subset pairs; no p-value is attached, because the cells are not independent replications—see the text). The low R^2 over the fitted range is itself evidence that d-support is a weak driver. Some late-life cells also show a sizeable *raw* transverse response, which is why the line fitted through them is a conservative comparator; those cells nonetheless select $\mathcal{U}_\parallel$ rather than $\mathcal{U}_\perp$, because the decision uses control-normalised responses and $\mathcal{U}_\parallel$ moves far more there (Tables 10 and 11). Dose values are in Table 12.

G Synthetic Ground-Truth Intervention

The construct interventions of Appendix F perturb a reference distribution whose true generative factors are unknown. This appendix runs the complementary check on data whose factors we set: it asks whether, when three factors are manipulated by construction, the intended PCNG coordinate responds. It is a positive control on the readout, not a validation of the coordinates on real data, and we read it only that way.

G.1 Protocol

Three latent factors—progression s, transverse offset r, and traversal rate —generate observations and interventions and are *never* supplied to the encoder or to PCNG:

$$x_t = f(s_t, r_t) + \epsilon_t, \qquad s_{t+1} = s_t + \dot{s}\,\Delta t. \tag{12}$$

Rate enters only through the state update, so $\dot{s}$ is recoverable from the temporal structure of a window and not from any single observation; an earlier design that placed $\dot{s}$ inside f made the rate result circular and is discarded. The encoder is an unsupervised reconstruction autoencoder with a two-dimensional bottleneck fitted on normal observations only. PCNG is then applied exactly as on real data.

G.2 Result

Table 13 reports the mechanism-classification outcome.

Rate is a positive control, and it passes. With $\dot{s}$ removed from f, failure was possible: the rate coordinate could only fire if traversal speed left a recoverable trace in the window's temporal structure. It fires in every repetition of every seed. This does not overturn the weak real-data rate result of Appendix F—the two are different questions—but it does separate them. The scoring rule for $\mathcal{U}_{\dot{s}}$ is not inert; what the real packs do not supply is a representation in which traversal rate survives.

Table 13: **Mechanism classification under known interventions.** A cell counts repetitions in which the intended component has the largest response; 12 repetitions per seed, five seeds.

Manipulated factor	Intended component	Selected
Transverse offset r	$\mathcal{U}_\perp$	$\mathbf{5 \times 12/12}$
Traversal rate	$\mathcal{U}_{\dot{s}}$	$\mathbf{5 \times 12/12}$
Progression s	$\mathcal{U}_\parallel$	$\{0, 1, 2, 8, 9\}/12$ —*not claimed*

Construct validity is representation- and data-dependent, and we do not manufacture a synthetic rate failure to match the real one.

Progression fails here, and that is informative. The progression manipulation selects $\mathcal{U}_\parallel$ in 20 of 60 repetitions, which is chance for a three-way argmax, and the per-seed counts are $\{0, 1, 2, 8, 9\}$ out of 12: two seeds select it most of the time and two almost never. We therefore claim no synthetic validation for it, and the seed spread rather than the pooled rate is the reason: an effect that reverses this completely across noise draws is not a measured construct, whatever its average. There is also a structural reason not to expect a stable one here. $\mathcal{U}_\parallel$ is a two-sided tail probability in s: moving a unit along a well-covered manifold changes its position but not the rarity of that position, so $\mathcal{U}_\parallel$ does not respond. It responds when a unit sits at an *unusual* progression relative to the reference. That is exactly why the late-life holdout in Appendix F scores $36/36$—the ablation removes the range of s the probe occupies, making those positions rare by construction—and it is the evidence behind our describing that row as close to definitional rather than as support for a progression interpretation.

Scope. One generative process, one encoder architecture, five seeds varying the noise draw rather than the generative assumption. A coordinate can pass on a favourable synthetic process and still fail on real packs, and one does. Geometric availability is not semantic validity; that is the distinction this appendix and Appendix F exist to keep separate.

H Bootstrap Stability of the Measured Geometry

A measurement protocol should say how much its readings move when the reference sample changes. This appendix resamples the reference population and reports what survives.

H.1 Protocol

For each of 12 matched CNN jobs (DS01–DS04, seeds 42–44) we draw 500 bootstrap resamples of the training *units* with replacement—units, never windows, since windows within a unit are not exchangeable—refit the principal curve on each resampled reference population, and reproject the *same* held-out test windows. The construct decision is recomputed on the 12 directed subset pairs at seed 42 against every resample.

H.2 What is stable

Table 14: **Geometry under reference resampling**, 500 resamples per job. Correlations are between resampled and original coordinates on identical held-out windows; the curve distance is normalised by curve length.

Quantity	median	5th pct. (95th for distance)
$\rho(s^{(b)}, s^{(0)})$	0.934	0.820
$\rho(d^{(b)}, d^{(0)})$	0.759	0.621
curve–curve distance	0.046	0.065
$\Pr$(same construct winner)	0.838 (all pairs)	0.890 (pairs whose baseline winner was $\mathcal{U}_\perp$)
$\Pr$(same gate label)		0.149

Call a resample *decision-consistent* when the cross-subset manipulation selects the same component it selected on the unperturbed reference. Across 500 resampled reference populations the decision is

consistent in 83.8% of cases overall, and in 89.0% of the pairs whose unperturbed winner was $\mathcal{U}_\perp$. Arc length is stable ($\rho = 0.934$ median, 0.820 at the 5th percentile) and the refitted curves stay close to the original.

H.3 Three stability regimes, not one

The useful finding is that these quantities do not share a stability regime, and testing them separately is what reveals it. Raw geometric coordinates are moderately stable; the construct *decision* built on them is more stable still; the categorical gate label is the least stable of the three. An argmax is not inherently robust—it is robust exactly when the winning margin is wide, and flips readily when two components are close—so decision stability is an empirical property to be measured, which is what the figure above does, not a consequence of taking a maximum.

The transverse coordinate moves most. $\rho(d^{(b)}, d^{(0)})$ has median 0.759 and a 5th percentile of 0.621, markedly below arc length, and DS04 is the weakest cell on both $\rho(d)$ and curve distance. $\mathcal{U}_\perp$ is built on d, so the coordinate the paper leans on hardest is the one that moves most—which is precisely why the decision-level figure above, rather than the coordinate correlation, is the number we quote for reliability.

The ternary gate label is threshold-brittle. $\Pr(\text{same GO/MARGINAL/NO-GO label})$ is 0.149, and every one of the twelve jobs falls between 0.050 and 0.286. This is not comparable to a $1/3$ guessing rate: each resample is scored against *that run's own* baseline label rather than against a uniform draw, so a run whose point estimate sits beside a cut can disagree with itself almost surely. Most of these jobs do sit near the $\mathrm{CV} = 0.15$ or 0.35 cut, and unit resampling moves the coefficient of variation across it. The Capacitor result illustrates that instability directly: post-hoc $\mathrm{CV}(L)$ is $0.377/0.190/0.368$, straddling the NO-GO/MARGINAL boundary across seeds. The gate should therefore be reported as a CV *distribution* with seed-level labels rather than as a categorical per-dataset verdict on borderline data.

Curve orientation is not fixed by the fit. A few runs return a negative median $\rho(s)$ because the refitted curve reverses direction. Magnitudes and distances remain informative, and signed correlations alone overstate instability; a production implementation should orient the curve by the within-unit time index.

I Residual-Scramble Falsification of the Transverse Construct Result

Section 5.3 reports that a cross-subset failure-mode shift raises $\mathcal{U}_\perp$ in 35 of 36 cells. That result is observational: it shows the component responds, not that the response is carried by the structure the component is supposed to measure. Here we destroy that structure and ask whether the result dies.

I.1 The perturbation

Using the existing fitted curve, each window is decomposed into its foot point $\hat{z}_i = \mathrm{proj}_\gamma(z_i)$ and transverse residual $r_i = z_i - \hat{z}_i$. Windows are binned by s under the paper's own rule ($n_{\mathrm{bins}} = \min(12, \max(3, n/30))$), residual *vectors* are permuted within a bin, and the latent is reconstructed as $z' = \hat{z}_i + r_{\pi(i)}$ before d and then $\mathcal{U}_\perp$ are recomputed through the unchanged rarity protocol. Permuting the vector rather than the scalar distance keeps z' in the form every real point has. Three conditions, 20 permutation draws each:

- D_0: no permutation.
- D_{cross}: permute across units within the bin, deranged at unit level. Destroys *which unit* carries the transverse residual.
- D_{within}: permute inside the same unit within the bin. Destroys fine temporal alignment while preserving unit identity.

s is preserved exactly by construction. The marginal distribution of d is *not* guaranteed to be, because a residual orthogonal to the curve at its old foot point is not exactly orthogonal at its new one; we therefore check it rather than assert it. The median within-bin Kolmogorov–Smirnov statistic between d' and d is 0.089 (cross) and 0.073 (within), small enough that marginal-d confounding is limited but not zero.

I.2 Result

Table 15: **Construct response and alarm behaviour under residual scrambling.** Hit rates are over the 36 directed cross-subset cells; false-alarm rate is over the 12 matched near-failure jobs. Chance for a three-way argmax is $1/3$.

Condition	$\mathcal{U}_\perp$ hit rate	$\mathcal{U}_\parallel$ hit rate	KS(d', d)	near-failure false alarm
D_0	$\mathbf{0.972}$	1.000	0.000	$\mathbf{0.049}$
D_{within}	1.000	1.000	0.073	0.117
D_{cross}	$\mathbf{0.667}$	0.956	0.089	0.151

Table 15 collects the outcome, and the registered predictions behaved as stated. $\mathcal{U}_\parallel$ itself is unchanged, since s is preserved exactly, and its hit rate moves only from 1.000 to 0.956: that residual movement is not a change in $\mathcal{U}_\parallel$ but a consequence of the hit rate being a three-way argmax, which can flip when a competing component is perturbed. The cross-unit condition drops the $\mathcal{U}_\perp$ hit rate from 0.972 to 0.667; paired per cell this is $+16/{=}19/{-}1$. The 36 cells are 12 directed subset pairs by three seeds rather than 36 independent trials, so we quote the sign count and the effect size and not a p-value; the drop from 0.972 to 0.667 is what carries the result.

I.3 What this establishes, stated precisely

The transverse construct response depends *partly* on structured, unit-level transverse geometry. Three qualifications keep that from being overstated.

The response is reduced, not abolished. The drop recovers 48% of the distance from D_0 to chance and D_{cross} remains twice chance, so a real part of the $35/36$ lives in the stage-conditioned marginal of d rather than in unit-level structure.

The effect is concentrated and uneven. Nineteen of the 36 cells do not move at all and the median paired difference is zero; the drop comes from the 16 cells that degrade, and it is heterogeneous among them: the mean decision loss among degraded cells is 0.691, with three collapsing completely ($\Delta = 1.00$), nine losing most of their decisions ($0.5 \le \Delta < 1.0$), and four barely moving. The sole improver gains 0.05. Because cells flip *partially* across the 20 draws, the 16 degraded cells sum to 11.05 cell-equivalents rather than 16: the aggregate $35/36 \to 24/36$ is a loss of 11 cell-equivalents, not 11 cells flipping outright. Cells with a wide argmax margin survive scrambling, which is expected and is not evidence of anything further.

The perturbation scale is one cell. D_{within} returns 1.000 against the unperturbed 0.972, and the difference is exactly one of the 36 cells—the same DS04 $\to$ DS02, seed-44 cell that is D_0's sole miss, moving across the decision boundary rather than a new population of cells becoming hits. That cell is borderline in every condition: it also supplies the single -1 in the paired sign count, improving by 0.05 under cross-unit scrambling. One cell is therefore the scale at which a structure-preserving perturbation moves this measurement, and the cross-unit effect of eleven cell-equivalents is measured against it. We state the anomaly rather than leave a reader to notice that a perturbation improved a score.

An asymmetry between the two readouts. Within-unit scrambling leaves the construct argmax perfect (1.000) while more than doubling the near-failure false-alarm rate ($0.049 \to 0.117$). The alarm therefore depends on fine temporal alignment inside a unit; the construct argmax does not. The two experiments read different properties of the same coordinate, and only the alarm is sensitive to when within a unit the transverse deviation occurs.

What it does not establish. Nothing here bears on whether $\mathcal{U}_\perp$ outperforms Mahalanobis distance—the tie in Section 8.2 stands—nor on component independence, nor on prediction error.

Verification of the unperturbed arm. A falsification test is interpretable only when its unperturbed control reproduces, under the same protocol, the quantity being falsified. We therefore verified D_0 before applying any perturbation: the implementation used above returns $35/36$ on the cross-subset construct and a near-failure false-alarm rate of 0.049, matching Sections 5.3 and 8.2 exactly, including the identity of the single missed cell.

An earlier implementation did not. It returned $32/36$ and a false-alarm rate of 0.076, through protocol drift—a random window drop in place of the block-unit control, a coarser curve fit, and refitted rather than stored coordinates—together with a difference in random-stream state at the point the control holdout is drawn. That last point is the reproducibility lesson worth recording: fixing a seed is not sufficient when the control is sampled from a generator that other steps have already advanced, so the reproducible object is the seed together with the number and order of draws, not the seed alone. Only the protocol above is reported in this paper.

Table 16: **Cross-domain residual coupling and admissibility.** "test units" labels the held-out unit count; the admissibility gate is computed on training units (PRONOSTIA/XJTU-SY: 12, Battery: 3, Capacitor: 8) using the same post-hoc principal curve that defines PCNG. Unit-preserving references cover PRONOSTIA and XJTU-SY. NO-GO rows are retained as reference-scaling diagnostics, not licensed semantic attribution.

Dataset	runs	test units	cond. $\max \lvert\rho\rvert$	indep. ref.	unit ref.	Gate
N-CMAPSS DS01–DS03	117	3–6	0.096 ± 0.054	0.025–0.041	0.049 *(9 runs)*	mixed by encoder
PRONOSTIA	3	5	0.298 ± 0.112	0.041–0.043	**0.254**	NO-GO $\times 3$
XJTU-SY	3	3	0.192 ± 0.004	0.055–0.077	**0.167**	NO-GO $\times 3$
NASA Battery	3	1	0.127 ± 0.039	0.121	—	MARGINAL $\times 3$
NASA Capacitor	3	3	0.249 ± 0.096	0.271	—	NO-GO/MARGINAL/NO-GO

J Correlation Magnitudes Under Both Novelty Calibrations

Each geometric coordinate is mapped to a novelty score before the components can be compared. The final paper uses tail rarity throughout. We retain the earlier density-surprisal definition for methodological context, but report all unit-dependent numbers under tail rarity only.

Density surprisal (earlier). A Gaussian kernel density estimate $\hat{p}_R^{(c)}$ is fitted on the reference values of coordinate c, and $\mathcal{U}_c = \min(-\log \hat{p}_R^{(c)}(c_i),\ \log N_R)$. The cap is required: a probe point outside the kernel's support otherwise reports an arbitrarily large value measuring extrapolation distance rather than novelty, and in one early run produced a response of 33 nats.

Tail rarity (used here). Each coordinate is first expressed as a raw score ($R_\perp = d^2/V_\parallel$, $R_\parallel = s$, $R_{\dot{s}} = \dot{s}$), and its empirical reference CDF F gives a tail probability—$q_\perp = 1 - F$ for the transverse score, which is novel only when large, and $q_c = 2\min(F, 1-F)$ for the two coordinates that are novel at either extreme. Then $\mathcal{U}_c = -\log \max(q_c, 1/N_R)$.

We adopted tail rarity for three reasons, all independent of its effect on our numbers. It has no bandwidth or any other free parameter, so the frame cannot be tuned. It is invariant to any monotone reparameterisation of a coordinate, which removes the question of why $d^2/V_\parallel$ rather than d. And the floor at $1/N_R$ is intrinsic rather than a bolted-on cap: the score saturates at the resolution of the reference sample by construction. Its one cost is that a two-sided tail probability treats a value between two modes of a multimodal reference as typical, where a density estimate would flag it; this is a real limitation for $\mathcal{U}_{\dot{s}}$ under heterogeneous operating conditions, and one reason we do not construct-validate $\mathcal{U}_{\dot{s}}$.

Current numerical comparison. Under the released tail-rarity pipeline, the 117-run raw maximum is 0.259 ± 0.118 and the conditional residual-coupling maximum is 0.096 ± 0.054; the construct results remain $35/36$ for the transverse intervention, $36/36$ for the late-life calibration check, and $21/36$ for the rate intervention. The rate construct therefore remains unresolved. The admissibility gate is computed before rarity calibration, so the calibration choice cannot improve an admissibility verdict.

K Conditional-Correlation Rescored Detail

The final $R_c \to U_c$ pipeline was rescored without retraining encoders, principal curves, or RUL readouts. Across the 117 primary runs, the raw maximum pairwise correlation is 0.259 ± 0.118 and the conditional rank-residual maximum is 0.096 ± 0.054. The independent-row Monte Carlo remains descriptive only and is not used to claim independence.

L Correlation Audit and Progression-Coordinate Sensitivity

Table 19 records the audit summary behind the pooled residual-coupling statistic of Table 18. The complete 117-cell machine-readable table is generated by the repository audit script.

For the admissibility gate we lock one definition: the same training-only post-hoc principal curve that induces the PCNG $s, d, \dot{s}$ coordinates also supplies per-unit traversal length. Stored alternative progression coordinates are sensitivity checks, not alternative gates. Table 17 measures how much the choice matters. Every column uses the repository routine `step1_pathlength` (the admissibility calculation) and the same unit identities; only the progression coordinate differs.

Table 17: **Admissibility is progression-coordinate sensitive.** CV(L) per external domain (mean over three seeds) under the canonical training-only post-hoc curve and under three stored progression coordinates produced by the trained trajectory model. Unit identities and the path-length routine are identical across columns. The last column counts, out of three seeds, runs whose thresholded label under s_{train} differs from the label under the canonical coordinate; it is specific to that pair and is not a count over all stored coordinates. Dashes mark coordinates a domain's runs do not store.

Domain	post-hoc s	s_{train}	s_{deg}	s_{gated}	flips vs. s_{train}
PRONOSTIA	$\mathbf{0.961}$	0.324	0.516	0.456	$2/3$
XJTU-SY	$\mathbf{0.560}$	0.469	0.876	0.618	$0/3$
Capacitor	$\mathbf{0.312}$	0.362	—	—	$1/3$
Battery	$\mathbf{0.315}$	0.228	0.253	—	$0/3$

The sensitivity is material rather than incidental. PRONOSTIA moves from a mean CV(L) of 0.961 under the canonical coordinate to 0.324 under s_{train}, which is the difference between a decisive decline and two of three seeds reading MARGINAL; on seed 44 a stored gated progression gives 0.122, inside the GO band. Capacitor's majority verdict also depends on the coordinate. Across the twelve external runs the thresholded label differs between the canonical and s_{train} columns on three. The direction is not consistent across coordinates: s_{train} reads lower than the canonical curve on PRONOSTIA, XJTU-SY and Battery but higher on Capacitor, while the stored degradation and gated coordinates read *higher* than canonical on XJTU-SY. No stored coordinate is uniformly more permissive, so this is coordinate sensitivity rather than a single optimistic bias awaiting a fix.

We draw the conservative conclusion. The gate is not only threshold-brittle, it is progression-coordinate-brittle, and the remedy is to name one coordinate prospectively rather than to select per dataset. The canonical choice follows from what the paper already defines PCNG to be: the coordinate that gates admissibility is the same training-only post-hoc arc length that induces $s, d, \dot{s}$. Choosing s_{train} because it rescues PRONOSTIA would let a dataset pass the gate under one geometry and be interpreted under another, which is a harder position to defend than a stricter external result.

Table 18: **Conditional rank-residual correlation across 117 runs, descriptive.** Geometry is extracted identically for every configuration. The statistic is the largest Spearman correlation among the three rarity components after conditional rank residualisation on s; larger values indicate stronger residual coupling. The lower block decomposes the maximum by pair, "argmax" counting runs in which that pair is largest (columns sum to 117). Unit-preserving references cover 9 MonoCNN development runs plus 6 external runs (Appendix E). This table is descriptive and does not support an independence claim; it is reported here rather than in the main text.

Quantity	Value	
Raw $\max \lvert\rho\rvert$	0.259 ± 0.118	
Conditional rank-residual $\max \lvert\rho\rvert$	0.096 ± 0.054	
Runs / frozen encoders	$117/13$	DS01–DS03, seeds 42–44
By pair, after residualisation:		
$\mathcal{U}_\perp$–$\mathcal{U}_\parallel$	0.070 ± 0.049	argmax $57/117$
$\mathcal{U}_\perp$–$\mathcal{U}_{\dot{s}}$	0.057 ± 0.057	argmax $40/117$
$\mathcal{U}_\parallel$–$\mathcal{U}_{\dot{s}}$	0.035 ± 0.026	argmax $20/117$
Unit-preserving reference, 15 *recomputed runs (Appendix E):*		
observed	0.056–0.425	
reference $p_{97.5}$	0.084–0.504	$4/15$ above

Table 19: **Correlation audit summary.** Machine-readable outputs cover 117 runs (pooled conditional $\max \lvert\rho\rvert$ from the released rarity pipeline). The MonoCNN row matches the nine observed values in Table 8 exactly.

Audit quantity	Value	Note
All 117: conditional $\max \lvert\rho\rvert$	0.096 ± 0.054	pooled output
MonoCNN 9: conditional $\max \lvert\rho\rvert$	0.109 ± 0.039	range 0.056–0.169

M Admissibility Under Controlled Geometric Stress

The mixed Capacitor gate outcome in Section 7 motivates a broader question: as a structural assumption is violated, does the admissibility gate withdraw admissibility before the component interpretations stop being reliable? The controlled construction below answers partly yes and partly no, and the "no" delimits where the gate can be trusted.

M.1 Construction

A curve $\gamma(s)$ in $\mathbb{R}^8$, $s \in [0, 1]$; 24 unit trajectories of 60 windows, $z = \gamma(s) + \sigma\varepsilon$ with isotropic Gaussian ε and per-unit rate from $U(0.8, 1.2)$; half the units are the reference set, half the probe set; 15 seeds per cell. Four stress axes are swept one at a time, each with a null level reproducing the assumed geometry. At each cell we record the admissibility decision (GO when the coefficient of variation of traversed per-unit path length is below 0.15) and a construct-accuracy score: three known coordinate perturbations are applied to a fresh probe population and we check whether the intended component's control-normalised response is the largest of the three.

Calibrating the perturbations. An earlier version of this sweep scaled the perturbations to the stress parameter ($d \to d + 6\sigma$, $\dot{s} \to 3\dot{s}$). Construct accuracy was then 1.00 in every cell, including fully separated branches: the test had no power, because the perturbation grew with the difficulty. We therefore use *absolute* offsets $(s + \Delta s,\ d + \Delta d,\ \dot{s} + \Delta\dot{s})$, calibrated once at the null stress level to the smallest offset at which the intended component is selected in at least 80% of seeds, and frozen for the whole sweep: $\Delta s = 0.02$, $\Delta d = 0.01$, $\Delta\dot{s} = 0.005$. Construct accuracy now moves under stress, which is what makes the rest of this appendix interpretable.

M.2 Result

Three readings, and we hold all three.

Table 20: **Gate withdrawal against component degradation.** Construct accuracy at the null level of each axis, its minimum over the swept range, and the crossing levels. Both crossing columns report the *first swept level* at which the condition holds, so the two are directly comparable: "first gate-out" is the first level with $P(\mathrm{GO}) < 0.5$, and "first -20% acc." the first at which construct accuracy has fallen a fifth below its null value. Because the grid is coarse, these overstate the crossing point; the linearly interpolated crossings are $\theta = 0.58$ for branching—the value quoted in the text—and a back-fraction of 0.32 for non-monotone, where the text instead quotes the first swept level, 0.4, at which $P(\mathrm{GO})$ has already reached 0.00. Component degradation is scored as a relative drop from each axis's own null level, which is stricter than requiring an absolute failure—no axis ever fails absolutely within range. Chance is $1/3$; no cell approaches it.

Stress axis	null acc.	min acc.	first gate-out	first -20% acc.	which comes first
Branching (θ)	0.93	0.60	0.80	0.20	component degrades first
Curvature (κL)	0.98	0.78	never	7	component degrades, gate blind
Transverse noise	1.00	0.82	never	never	neither crosses
Non-monotone (back-fraction)	0.98	0.76	0.40	0.50	gate withdraws first

The gate does withdraw where the model is wrong about paths. Under non-monotone progression it withdraws before component accuracy degrades appreciably ($P(\mathrm{GO}) = 0.00$ at a back-fraction of 0.4, where accuracy is still 1.00), and under branching it eventually withdraws on most seeds ($P(\mathrm{GO}) = 0.20$ at $\theta = 0.8$). Both are geometries that change the distribution of traversed per-unit path length, which is exactly what the CV statistic measures. The elevated Capacitor CVs are consistent with this family, but its seed-level labels straddle a threshold, so we do not treat that dataset as a categorical stress-case demonstration.

But the gate is blind to stressors that leave path length intact. Curvature and transverse noise never drive $P(\mathrm{GO})$ below 0.5—the minimum is 0.60 and 0.87 respectively—while construct accuracy falls from 0.98 to 0.78 and from 1.00 to 0.82. A curved or noisy manifold leaves every unit travelling a similar distance, so a path-length statistic cannot see it. The admissibility gate is a test of *path homogeneity*, not of geometric reliability in general, and we say so rather than implying broader coverage.

And on branching, degradation precedes withdrawal. Construct accuracy has already fallen from 0.93 to 0.73 at $\theta = 0.2$ and to 0.60 at $\theta = 0.35$, while $P(\mathrm{GO})$ is still 0.87 and 0.67. Interpolating between the swept levels $\theta = 0.55$ ($P(\mathrm{GO}) = 0.53$) and $\theta = 0.8$ (0.20), the gate crosses 0.5 only near $\theta = 0.58$; the first swept level below 0.5 is $\theta = 0.8$. So on the axis that most directly violates the single-curve assumption, the components become less reliable before the gate objects. The degradation is graceful—nothing approaches the one-in-three chance level anywhere in the sweep—but it is not detected early, and a user of this instrument should not read GO as a guarantee that the component interpretations are intact.

Reading the numbers. Neither curve is monotone at 15 seeds: $P(\mathrm{GO})$ on the branching axis is 0.20 at $\theta = 0.8$ and 0.47 at $\theta = 1.1$, and construct accuracy rises again from 0.60 to 0.73 over the same interval. We therefore quote crossing levels as approximate operating points of one synthetic construction and draw no general thresholds from them. What survives the noise is the pattern in Table 20: the gate covers one family of failures and not the other, and it is not an early warning even within the family it covers.

M.3 Admissibility across encoder groups

The gate is a property of the geometry a representation exposes, not of the dataset alone, so it should vary across encoders trained with different objectives on the same data. It does. Table 21 reports admissibility over the 117 development runs—DS01–DS03 by three seeds by the 13 frozen configurations of Section 4—grouped by encoder family. (DS04 enters the onset and near-failure sweeps of Appendices O and R, not this gate table.)

Three things follow, and one of them is inconvenient.

Table 21: **PCNG admissibility across encoder groups**. A group is one encoder architecture as listed in Section 4; n counts runs (dataset $\times$ seed $\times$ latent dimension), so the effective number of independent representations per row is one. GO rate is the fraction with CV$(L) < 0.15$; the last column is the residual-coupling statistic of Table 18.

Encoder group	n	Gate majority	GO rate	mean CV(L)	mean $\max \lvert\rho\rvert$
Tr.AE	9	GO	1.00	0.101	0.114
MonoTR	9	GO	0.89	0.108	0.115
STEP	9	GO	0.78	0.110	0.121
BiLSTM-AE	9	GO	0.78	0.102	0.072
SoftCLT-TS2Vec	27	GO	0.52	0.155	0.095
MonoCNN	9	MARGINAL	0.44	0.163	0.109
TS2Vec	36	MARGINAL	0.36	0.176	0.080
SoftCLT-TR	9	MARGINAL	0.33	0.165	0.114

Admissibility is representation-dependent. GO rates span 0.33 to 1.00 on identical data. The same fleet, the same windows and the same gate give different verdicts depending only on how the encoder was trained. That is the empirical content of the scope statement in Section 3.2: PCNG reads progression geometry, it does not create it, and whether a representation exposes it is a property of the training objective.

Admissibility does not predict residual coupling. The two columns answer different questions. Tr.AE admits every run and reaches $\max |\rho| = 0.114$; SoftCLT-TR admits a third and reaches the same 0.114; BiLSTM-AE admits 78% of runs with the lowest residual dependence (0.072). A representation can be cleanly traversable and still carry residual coupling, or the reverse. The gate asks whether one progression curve is a defensible summary; the correlation diagnostic asks whether the calibrated coordinates share residual structure.

Our own primary encoder is among the less admissible groups. MonoCNN—the encoder behind the construct results of Appendix F and the alarm of Section 8.2—has mean CV$(L) = 0.163$ and a GO rate of 0.44, giving a MARGINAL majority. Its mean CV is the third highest among the eight encoder groups. We report the downstream results because MARGINAL is a proceed-and-disclose verdict under our protocol; the gate therefore does not privilege the contributed encoder.

What this table does not support. Each row is one encoder architecture observed over datasets, seeds and latent dimensions, so n counts runs while the number of independent representations per row is one. The comparisons are therefore descriptive: we do not claim that contrastive objectives are less admissible than reconstruction objectives. In this panel GO rates span 0.33–1.00 and mean CV(L) spans 0.101–0.176; establishing an objective-level effect would require several independently trained encoders per objective. Construct validity was measured for MonoCNN alone, so no column of this table should be read as comparing construct validity across groups.

M.4 An amended-preregistration controlled construct-level transfer outside prognostics

A separate controlled CIFAR-10 corruption-transfer construction asks whether the *attribution*, rather than the downstream alarm, transfers outside prognostics. Its pre-registration is an *amended* one, and we describe the amendment rather than present it as a blind lock. A first version was run and produced output, which we voided on inspection for three implementation defects: the progression measure pooled all images instead of taking a per-image Spearman correlation and averaging over images, the family-shift measure compared unpaired group means instead of differences matched on image identity and severity, and the severity-zero state was produced by calling the corruption library rather than by using the original image. Version 2.1 (`CIFAR_VISION_TRANSFER_PREREG.md`) repairs those three points and also revises the PASS criterion, so we state the revision rather than present it as a restatement. The directional hypothesis is unchanged—family shift toward $\mathcal{U}_\perp$, severity toward $\mathcal{U}_\parallel$—as is the requirement that both family cells hit on $\geq 2/3$ seeds. What changes is the progression bar, which moves with the statistic it is written against, from a pooled

$\rho(s, \text{severity}) \geq 0.3$ to a per-image mean $\rho \geq 0.5$ on $\geq 2/3$ seeds; and the severity-only cell, which was a v1 PASS requirement and is descriptive in v2.1. All of this was written after the voided output had been seen, so this is an amended, non-blind pre-registration and we claim nothing stronger. The revision does not manufacture the verdict: the reported run also satisfies the v1 criterion, with pooled ρ of 0.949, 0.955 and 0.936 against its 0.3 bar and the severity cell selecting $\mathcal{U}_{\|}$ on all three seeds. The void record (`CIFAR_CONSTRUCT_VOID.md`) and both pre-registration documents ship with the supplement, so the two criteria can be diffed directly. The script is `scripts/run_pcng_cifar_construct.py`. It uses CIFAR-10 training images with the `imagecorruptions` implementation of CIFAR-10-C corruption-family names and severity levels 1–5, plus the original image as severity 0; it does not load the official CIFAR-10-C `.npy` arrays. It is neither the paper-style reference-ablation intervention nor the pre-registered fleet-gate experiment below. Corruption severity supplies a known progression factor and held-out corruption family supplies a severity-matched transverse change. The reference is image half A under `gaussian_noise`; image half B is held out and scored under the same gaussian control plus `motion_blur` and `contrast`, always on the same severity grid $\{0, 1, 2, 3, 4, 5\}$. The curve and rarity calibration are fit once on the reference and never on the probes. The reported $\text{CV}(L)$ is a construction diagnostic only and is never an admissibility decision.

For the severity-matched family-shift test, the response is paired by image and severity,

$$r_k = \mathbb{E}_{(i,\text{sev})}[U_k(\text{family}; i, \text{sev}) - U_k(\text{gaussian}; i, \text{sev})]\,, \tag{13}$$

and is normalised by a between-image gaussian negative control,

$$\tilde{r}_k = \frac{r_k}{\max\{|r_k^{\text{ctrl}}|, \max(10^{-6}, 0.25\,\text{mean}_j|r_j^{\text{ctrl}}|)\}}. \tag{14}$$

The negative control is a normaliser only, not the matched pair itself.

The locked decision grid first terminates as **FAIL (projection sanity)** if $|\rho(\text{PC}_1, \text{severity})| < 0.1$ on the reference split on at least $2/3$ seeds, with PC_1 denoting the first principal-component score. Otherwise, **TRANSFER PASS** requires (i) mean within-image $\rho(s, \text{severity}) \geq 0.5$ on probe-gaussian trajectories on at least $2/3$ seeds, (ii) both severity-matched family probes to be directional hits for $\mathcal{U}_{\perp}$ on at least $2/3$ seeds, and (iii) the locked held-out image split. A family cell is a directional hit when $\arg\max_k \tilde{r}_k = \mathcal{U}_{\perp}$ and $\tilde{r}_{\perp} > 0$, where $\tilde{r}_k$ is the pre-specified control-normalised paired response over image$\times$severity states. The residual-coupling diagnostic and the severity-only arm are report-only and never enter the decision.

The observed outcome satisfies that grid and is therefore labelled **TRANSFER PASS** by the pre-registration; the label is provenance, not a significance claim. The mean within-image Spearman correlation $\rho(s, \text{severity})$ is 0.973 ± 0.044, 0.978 ± 0.033, and 0.962 ± 0.075 on seeds 42–44, with fractions of images above 0.5 equal to 1.00, 1.00, and 0.99. For the paired, severity-matched family shift, both held-out families select $\mathcal{U}_{\perp}$ on seeds 42 and 44, whereas both select $\mathcal{U}_{\|}$ on seed 43: the directional hit is therefore $4/6$ family cells, or $2/3$ seeds. The severity-only $\mathcal{U}_{\|}$ argmax holds on all three seeds but is descriptive rather than part of the PASS criterion. The residual-coupling statistic is also descriptive only and supports no independence claim. With two families and three seeds we make no statistical-significance claim. We use this experiment only as a controlled construct-level transfer check, not as a pre-registered fleet-gate result, a reference-ablation result, a vision-OOD benchmark, or evidence that PCNG generally transfers to vision.

This amended-preregistration transfer construction is separate from the fleet-gate experiment below, which was pre-registered blind; the two use different constructions and answer different questions.

M.5 A PRE-REGISTERED REFUSAL OUTSIDE THE DEVELOPMENT DOMAIN

The mixed Capacitor gate outcome of Section 7 arose during development. Separately from the controlled transfer construction above, we tested whether the gate behaves the same way on a domain chosen in advance and far from prognostics. We pre-registered an application to CIFAR-10-C before any number from that construction existed: the trajectory design, the reference and probe corruption splits, the gate thresholds, the sample size, the projection diagnostic and the reading for every outcome were fixed and timestamped first. We report it because a pre-registration reported only when it passes is worth nothing.

Construction. A frozen torchvision ResNet-18 with ImageNet-1K weights encodes CIFAR-10-C images: the classification head is replaced by an identity map, giving a 512-d feature, and the network is used in evaluation mode with no fine-tuning on CIFAR. Images are upsampled from 32×32 to 224×224 and normalised with ImageNet statistics. The feature is projected to 32 dimensions on the reference population only. One base image under one corruption type is a *unit*, and its trajectory is that image rendered along a severity schedule, so the step index supplies the clock that $\dot{s}$ requires. Three fixed-length schedules—$(1, 2, 3)$, $(1, 2, 4)$ and $(1, 3, 5)$—give equal window counts with three different traversal extents, so per-unit path length varies for reasons the gate is meant to measure rather than through sampling density. Eight corruption types form the reference population and seven are held out as probes. The admissibility gate runs first on the reference units, and the protocol stops there if it declines.

Result: NO-GO on every seed. Table 22 reports the gate outcome.

Table 22: **Pre-registered admissibility outcome on CIFAR-10-C.** $\rho(\mathrm{PC}_1, \mathrm{severity})$ is the pre-registered projection diagnostic, with PC_1 denoting the first principal-component score; values above 0.1 mean the projection retained a severity signal, so a decline cannot be attributed to the projection discarding it.

Seed	$\rho(\mathrm{PC}_1, \mathrm{sev})$	cum. variance	$\mathrm{CV}(L)$	Verdict
42	0.152	0.731	1.296	NO-GO
43	0.158	0.729	1.179	NO-GO
44	0.154	0.727	1.114	NO-GO
Pooled			$\mathbf{1.196 \pm 0.092}$	NO-GO

The gate declines on all three seeds at a coefficient of variation of 1.196 ± 0.092, roughly three and a half times the NO-GO threshold of 0.35. This is not a marginal call. The pre-registered projection diagnostic clears on every seed, so the decline is not an artifact of the projection discarding severity, and no rarity score was computed.

What we conclude, and what we do not. The admissibility gate declines CIFAR-10-C under this construction. Whether that generalises to vision is a second question this run does not answer, and we do not answer it. We note only that the severity signal surviving into the projection is weak ($\rho \approx 0.15$), and that two features of this encoder work against it. An ImageNet classification objective rewards *invariance* to appearance perturbation, so corruption severity is close to the kind of variation such a representation is trained to discard; and upsampling 32×32 inputs to 224×224 attenuates the high-frequency signatures that several of the reference corruptions consist of. This encoder is therefore not a neutral choice but an unfavourable one, which is the point: it is the generic frozen representation a practitioner would reach for first, and Section 3.2 already warns that our own encoders carry a monotone-degradation prior from the temporal index rather than being assumption-free. Separating an encoder property from a domain property would require a representation trained to retain severity, which we have not run. What the result does establish is that the gate is not ornamental: applied to a domain chosen in advance, with the reading fixed beforehand, it refuses, and the protocol stops.

Earlier construction. An initial design held traversal extent equal across all units, which suppresses exactly the between-unit path-length variation the admissibility gate measures and is therefore biased toward admitting. It returned $\mathrm{CV}(L) = 0.854$, also a decline. We report it because it was obtained under conditions favourable to passing; the pre-registered construction above supersedes it.

N Detailed Onset-Detection Results

A compact onset-detection table from an earlier development version is retained here for continuity; Table 24 is the authoritative sweep, and Table 31 the headline summary.

Table 23: **Onset detection under two independent detectors.** Scores are binarised at the 95th training pre-onset percentile and integrated by either a Bernoulli Page CUSUM (Page, 1954) or a causal rolling spike rate (K=20), so the two blocks are matched across methods; baselines use a Mahalanobis score (Mahalanobis, 1936) on the frozen latent under the identical transform. Spike fold is the post/pre spike-rate ratio: a value near 1 means the alarm fires equally often before and after onset and discriminates nothing. Rows marked † fail that test and are excluded from the ranking. Bold and second-place marks are relative to the subset of configurations shown here, so they do not agree cell-for-cell with the full sweep in Table 24, which contains higher values in several columns and is the table to cite. Full sweep: Appendix O.

		Page CUSUM AUROC ↑					**Rolling spike rate** AUROC ↑					**Spike fold** ↑	
Method	**Score**	DS01	DS02	DS03	DS04	Cross	DS01	DS02	DS03	DS04	Cross	DS01–03	Cross
MonoCNN (d=4)	$\lvert h\rvert$	0.958	0.992	0.942	0.765	0.758	**0.983**	**0.992**	0.952	0.801	0.744	14.0	3.1
MonoCNN (d=4)	$\lvert s\rvert$	**0.959**	**0.993**	0.942	0.718	0.761	**0.983**	**0.992**	0.946	0.728	0.749	13.9	3.4
MonoCNN (d=16)	$\lvert h\rvert$	0.945	0.974	**0.944**	0.812	0.771	0.977	0.971	**0.960**	0.839	0.779	12.4	2.9
MonoCNN (d=4)	$\lvert\mathcal{U}_{\mathrm{geo}}\rvert$	0.824	0.868	0.840	0.750	**0.909**	0.834	0.871	0.865	0.752	0.707	13.4	1.4
MonoCNN (d=4)	z-Mah.	0.931	0.983	0.928	0.747	0.843	0.948	0.984	0.942	0.773	0.781	14.0	1.8
SoftCLT-TS2Vec (d=4)	z-Mah.	0.930	0.971	0.937	**0.843**	0.901	0.954	0.977	0.952	**0.890**	**0.843**	19.4	1.8
TS2Vec (d=32)	z-Mah.	0.883	0.943	0.892	**0.843**	0.863	0.906	0.957	0.914	0.794	0.790	10.5	1.7
Tr.AE (d=4)	z-Mah.	0.806	0.880	0.812	0.635	0.718	0.838	0.919	0.841	0.610	0.712	9.2	5.1
SoftCLT-TS2Vec (d=320)†	z-Mah.	0.975	0.980	0.989	0.995	0.978	0.609	0.689	0.636	0.633	0.634	1.1	1.0
TS2Vec (d=320)†	z-Mah.	0.938	0.973	0.895	0.995	0.993	0.742	0.576	0.819	0.579	0.574	1.2	1.0

O Full Onset-Detection Sweep

Table 24 gives the complete sweep behind Tables 31–4: every latent dimension and every score, under both detectors, for the staged CNN route and for all baselines. Both detectors share the same spike transform—each score is thresholded at the 95th percentile of its own training pre-onset distribution—so the only difference between the two blocks is how the resulting binary sequence is integrated: cumulative log-likelihood ratio with a training-calibrated alarm threshold for the CUSUM, or a causal K=20 moving average compared against $\mu + 3\sigma$ of the training pre-onset rate.

The sweep supports three statements in Section 8.1. The rolling-rate detector is uniformly stronger in-domain, beating CUSUM in $38/40$, $34/40$ and $38/40$ matched configurations on DS01–DS03, while CUSUM is stronger on the cross-subset split in $38/40$; the two are therefore complementary rather than ordered, and we report both. The spike-rate ratio separates the degenerate configurations from every other row by an order of magnitude: the two d=320 baselines sit at 1.06 and 1.23 while no other configuration falls below 3.7. And the staged coordinates are the most detector-stable scores in the benchmark, which is what makes their in-domain advantage robust to the choice of detector rather than an artifact of it.

Table 24: **Full onset-detection sweep.** Every backbone, latent dimension and score under both detectors of Table 31. $\text{Fold}_{1\text{-}3}$ is the mean post/pre spike-rate ratio on DS01–DS03. Rows marked † are degenerate and excluded from the ranking. Mean over three seeds.

		Page CUSUM AUROC ↑					**Rolling spike rate** AUROC ↑					**Spike fold** ↑	
Method	**Score**	DS01	DS02	DS03	DS04	Cross	DS01	DS02	DS03	DS04	Cross	DS01–03	Cross
MonoCNN (d=2)	$\sqrt{h^2+U^2}$	0.797	0.900	0.849	0.655	0.823	0.811	0.920	0.870	0.654	0.719	13.6	1.7
MonoCNN (d=4)	$\sqrt{h^2+U^2}$	0.849	0.905	0.844	0.749	0.903	0.858	0.908	0.872	0.754	0.733	16.4	1.5
MonoCNN (d=8)	$\sqrt{h^2+U^2}$	0.802	0.876	0.908	0.736	0.828	0.825	0.876	0.928	0.748	0.762	11.2	1.7
MonoCNN (d=16)	$\sqrt{h^2+U^2}$	0.813	0.918	0.914	0.767	0.830	0.835	0.940	0.943	0.794	0.734	13.8	1.6
MonoCNN (d=32)	$\sqrt{h^2+U^2}$	0.890	0.945	0.943	0.792	0.816	0.913	0.961	0.960	0.814	0.755	12.5	1.7
MonoCNN (d=2)	$\sqrt{s^2+U^2}$	0.794	0.890	0.847	0.656	0.821	0.808	0.915	0.867	0.655	0.714	13.6	1.7
MonoCNN (d=4)	$\sqrt{s^2+U^2}$	0.841	0.897	0.844	0.749	0.902	0.849	0.903	0.873	0.756	0.728	16.4	1.5
MonoCNN (d=8)	$\sqrt{s^2+U^2}$	0.799	0.883	0.914	0.759	0.825	0.823	0.894	0.931	0.784	0.762	11.1	1.7
MonoCNN (d=16)	$\sqrt{s^2+U^2}$	0.829	0.906	0.918	0.782	0.826	0.853	0.925	0.944	0.816	0.729	14.3	1.6
MonoCNN (d=32)	$\sqrt{s^2+U^2}$	0.897	0.957	0.944	0.797	0.818	0.924	0.968	0.961	0.822	0.756	12.7	1.7
MonoCNN (d=2)	$\lvert h\rvert$	0.953	0.989	**0.948**	0.647	0.800	0.978	0.989	0.957	0.651	0.794	13.1	5.2
MonoCNN (d=4)	$\lvert h\rvert$	0.958	0.992	0.942	0.765	0.758	0.983	**0.992**	0.952	0.801	0.744	14.0	3.1
MonoCNN (d=8)	$\lvert h\rvert$	0.945	0.964	0.943	0.783	0.763	0.969	0.972	0.959	0.809	0.753	12.2	2.5
MonoCNN (d=16)	$\lvert h\rvert$	0.945	0.974	0.944	0.812	0.771	0.977	0.971	0.960	0.839	0.779	12.4	2.9
MonoCNN (d=32)	$\lvert h\rvert$	0.937	0.956	**0.948**	0.801	0.777	0.969	0.964	**0.965**	0.836	0.763	13.0	4.7
MonoCNN (d=2)	$\lvert s\rvert$	0.954	0.988	0.947	0.585	0.794	**0.984**	0.991	0.957	0.568	0.785	12.9	5.9
MonoCNN (d=4)	$\lvert s\rvert$	**0.959**	**0.993**	0.942	0.718	0.761	0.983	**0.992**	0.946	0.728	0.749	13.9	3.4
MonoCNN (d=8)	$\lvert s\rvert$	0.940	0.979	0.944	0.762	0.759	0.962	0.977	0.958	0.782	0.746	11.7	2.8
MonoCNN (d=16)	$\lvert s\rvert$	0.944	0.963	0.943	0.768	0.762	0.973	0.971	0.957	0.782	0.780	14.0	5.9
MonoCNN (d=32)	$\lvert s\rvert$	0.938	0.917	0.947	0.752	0.757	0.976	0.939	0.961	0.752	0.751	12.2	10.5
MonoCNN (d=2)	$\lvert\mathcal{U}_{\text{geo}}\rvert$	0.791	0.889	0.845	0.658	0.845	0.803	0.909	0.861	0.655	0.695	11.1	1.5
MonoCNN (d=4)	$\lvert\mathcal{U}_{\text{geo}}\rvert$	0.824	0.868	0.840	0.750	**0.909**	0.834	0.871	0.865	0.752	0.707	13.4	1.4
MonoCNN (d=8)	$\lvert\mathcal{U}_{\text{geo}}\rvert$	0.783	0.830	0.876	0.735	0.852	0.809	0.844	0.890	0.755	0.722	10.7	1.5
MonoCNN (d=16)	$\lvert\mathcal{U}_{\text{geo}}\rvert$	0.785	0.871	0.878	0.760	0.857	0.799	0.893	0.908	0.783	0.698	13.2	1.5
MonoCNN (d=32)	$\lvert\mathcal{U}_{\text{geo}}\rvert$	0.838	0.920	0.904	0.787	0.860	0.868	0.938	0.930	0.813	0.730	11.6	1.5
MonoCNN (d=2)	z-Mah.	0.913	0.978	0.936	0.661	0.841	0.942	0.987	0.944	0.649	0.813	16.3	2.1
MonoCNN (d=4)	z-Mah.	0.931	0.983	0.928	0.747	0.843	0.948	0.984	0.942	0.773	0.781	14.0	1.8
MonoCNN (d=8)	z-Mah.	0.896	0.936	0.926	0.785	0.849	0.920	0.954	0.944	0.817	0.795	10.9	1.7
MonoCNN (d=16)	z-Mah.	0.874	0.923	0.929	0.805	0.829	0.905	0.940	0.955	0.839	0.744	9.7	1.5
MonoCNN (d=32)	z-Mah.	0.895	0.964	0.941	0.807	0.821	0.918	0.966	0.957	0.832	0.750	8.5	1.5
BiLSTM-AE (d=2)	z-Mah.	0.696	0.748	0.775	0.642	0.751	0.696	0.750	0.739	0.584	0.605	4.7	1.4
BiLSTM-AE (d=4)	z-Mah.	0.722	0.856	0.798	0.526	0.758	0.725	0.876	0.819	0.550	0.694	8.9	1.8
BiLSTM-AE (d=32)	z-Mah.	0.763	0.865	0.792	0.659	0.845	0.792	0.879	0.824	0.644	0.673	6.6	1.4
STEP (d=2)	z-Mah.	0.876	0.944	0.905	0.529	0.716	0.919	0.958	0.932	0.536	0.700	9.4	1.6
STEP (d=4)	z-Mah.	0.871	0.936	0.920	0.739	0.763	0.935	0.945	0.935	0.753	0.711	10.1	1.6
SimCLR-TS (d=32)	z-Mah.	0.796	0.851	0.813	0.751	0.886	0.807	0.860	0.823	0.732	0.685	6.1	1.3
SoftCLT-TR (d=2)	z-Mah.	0.885	0.951	0.936	0.711	0.728	0.914	0.957	0.954	0.657	0.699	10.5	1.6
SoftCLT-TR (d=4)	z-Mah.	0.913	0.952	0.939	0.747	0.734	0.949	0.959	0.960	0.766	0.714	10.1	1.6
SoftCLT-TS2Vec (d=2)	z-Mah.	0.914	0.977	0.930	0.788	0.901	0.923	0.980	0.934	0.811	**0.843**	15.0	1.9
SoftCLT-TS2Vec (d=4)	z-Mah.	0.930	0.971	0.937	**0.843**	0.901	0.954	0.977	0.952	**0.890**	**0.843**	19.4	1.8
TS2Vec (d=2)	z-Mah.	0.664	0.655	0.867	0.605	0.683	0.604	0.688	0.892	0.550	0.553	6.6	1.0
TS2Vec (d=4)	z-Mah.	0.815	0.861	0.853	0.704	0.682	0.831	0.892	0.879	0.657	0.671	10.3	1.8
TS2Vec (d=32)	z-Mah.	0.883	0.943	0.892	**0.843**	0.863	0.906	0.957	0.914	0.794	0.790	10.5	1.7
Tr.AE (d=2)	z-Mah.	0.665	0.870	0.672	0.527	0.698	0.670	0.895	0.638	0.547	0.636	3.8	1.7
Tr.AE (d=4)	z-Mah.	0.806	0.880	0.812	0.635	0.718	0.838	0.919	0.841	0.610	0.712	9.2	5.1
TS2Vec (d=320)†	z-Mah.	0.938	0.973	0.895	0.995	0.993	0.742	0.576	0.819	0.579	0.574	1.2	1.0
SoftCLT-TS2Vec (d=320)†	z-Mah.	0.975	0.980	0.989	0.995	0.978	0.609	0.689	0.636	0.633	0.634	1.1	1.0

P NEAR-FAILURE ALARM AGAINST CLASSICAL NOVELTY SCORES: PAIRED AND PER-HOLDOUT DETAIL

Table 4 pools 12 matched jobs. Because the same detector, threshold and frozen embeddings are used for every score, each job supports a paired comparison, which is the informative unit: the scores differ only in what they measure, not in how they are thresholded.

Table 25: **Paired differences against $\mathcal{U}_\perp$**, 12 matched jobs. $\Delta\text{FA} = \text{FA}(A) - \text{FA}(\mathcal{U}_\perp)$, so a positive sign is a job on which the baseline raises more false alarms. $\Delta\text{lead} = \text{lead}_{\text{first}}(A) - \text{lead}_{\text{first}}(\mathcal{U}_\perp)$, so a positive sign is a job on which the baseline alarms *earlier*, which is worse for a near-failure alarm. p is an exact sign test on the non-tied jobs, taken one-sided in the direction the counts fall; because that direction is read off the data rather than fixed in advance it is anti-conservative by a factor of two, which is one reason these values are reported for completeness rather than treated as significance tests. Where no direction is predicted—the timing comparison among the Mahalanobis ties—we quote the two-sided value instead and say so. See Section 8.2 on seeds within a dataset.

	ΔFA			$\Delta\text{lead}_{\text{first}}$		
$A - \mathcal{U}_\perp$	mean	$+/0/-$	p	mean	$+/0/-$	p
$\mathcal{U}_{\text{geo}}$	$+0.035$	$2/10/0$	0.25	$+2.2$	$3/1/8$	0.11
Mahalanobis	$+0.042$	$1/11/0$	0.50	$+49.3$	$8/0/4$	0.19
kNN distance	$+0.069$	$4/8/0$	0.06	-11.6	$2/0/10$	$0.02^{\ddagger}$
KDE NLL	$+0.118$	$4/8/0$	0.06	$+69.8$	$9/0/3$	0.07
$\lvert h\rvert$ control	$+0.243$	$7/5/0$	$\mathbf{0.008}$	$+128.7$	$12/0/0$	$\mathbf{0.0002}$

‡ The one comparison that favours a baseline: kNN distance alarms later than $\mathcal{U}_\perp$ on 10 of 12 jobs. It also covers only 0.847 ± 0.253 of near-failure windows against 0.979 ± 0.072 for $\mathcal{U}_\perp$, so the later alarm coincides with missed detections rather than a cleaner timing trade-off.

Table 26: **Per-holdout breakdown**, three seeds each. Coverage is coverage of the final 20 cycles. The pattern that survives disaggregation is the false-alarm ordering; timing differences are driven by DS02 and DS04.

Holdout	Score	Coverage	FA	$\text{lead}_{\text{first}}$
DS01	$\mathcal{U}_\perp$	1.000 ± 0.000	0.000 ± 0.000	78.5 ± 63.9
	Mahalanobis	1.000 ± 0.000	0.000 ± 0.000	127.2 ± 17.8
	$\lvert h\rvert$	1.000 ± 0.000	0.167 ± 0.144	245.5 ± 21.6
DS02	$\mathcal{U}_\perp$	1.000 ± 0.000	0.000 ± 0.000	96.0 ± 37.3
	Mahalanobis	1.000 ± 0.000	0.000 ± 0.000	172.0 ± 72.2
	$\lvert h\rvert$	1.000 ± 0.000	0.556 ± 0.509	300.0 ± 59.3
DS03	$\mathcal{U}_\perp$	1.000 ± 0.000	0.111 ± 0.096	206.3 ± 23.0
	Mahalanobis	1.000 ± 0.000	0.111 ± 0.096	190.7 ± 22.7
	$\lvert h\rvert$	1.000 ± 0.000	0.278 ± 0.096	289.0 ± 9.8
DS04	$\mathcal{U}_\perp$	0.917 ± 0.144	0.083 ± 0.144	76.2 ± 68.0
	Mahalanobis	0.917 ± 0.144	0.250 ± 0.433	164.5 ± 229.0
	$\lvert h\rvert$	0.917 ± 0.144	0.167 ± 0.289	137.2 ± 141.8

DS03 is where every score degrades together: false alarms rise for all six and first alarms move earlier, which is consistent with DS03 being the subset on which the construct ablations were also least stable (Appendix F). The table above lists three of the six scores per holdout; Table 28 carries the in-domain per-subset breakdown for the remaining baselines, and the joint degradation is read across the two. Degrading together is not the same as staying indistinguishable: DS04 is where the baselines separate most—Mahalanobis triples $\mathcal{U}_\perp$'s false-alarm rate and its timing variance explodes (± 229 cycles across three seeds), which is the behaviour one expects from a single-Gaussian model on the subset with the most heterogeneous operating profiles.

A note on what this experiment does and does not settle. All twelve jobs use the $z{=}2$ CNN arm. Two dimensions is the regime in which kernel density estimation and kNN distance are most reliable, so the classical baselines are being met on favourable ground, and $\mathcal{U}_\perp$'s advantage over KDE is correspondingly meaningful. The converse is that we cannot extrapolate: at $z{=}320$ the baselines would be estimating density in a space where they are known to degrade, and the comparison would need to be rerun rather than assumed.

P.1 What the Mahalanobis tie is made of

Table 25 records 11 exact false-alarm ties against Mahalanobis distance, and the main text reports that as no separation. The composition of those ties is worth stating, because it bounds what the comparison can show (Table 27).

Table 27: **Anatomy of the Mahalanobis false-alarm ties**, 12 matched jobs. Jaccard is computed over pre-onset alarm windows. $\Delta\text{lead} = \text{lead}_{\text{first}}(\text{Mahal}) - \text{lead}_{\text{first}}(\mathcal{U}_\perp)$ in cycles.

Job	FA $\mathcal{U}_\perp$	FA Mahal.	tie	Jaccard (pre)	Δlead
DS01 s42	0.000	0.000	yes	1.000	+98
DS01 s43	0.000	0.000	yes	1.000	+6
DS01 s44	0.000	0.000	yes	1.000	+42
DS02 s42	0.000	0.000	yes	1.000	+21
DS02 s43	0.000	0.000	yes	1.000	+136
DS02 s44	0.000	0.000	yes	1.000	+71
DS03 s42	0.167	0.167	yes	1.000	−0
DS03 s43	0.167	0.167	yes	1.000	−27
DS03 s44	0.000	0.000	yes	1.000	−20
DS04 s42	0.000	0.000	yes	1.000	+4
DS04 s43	0.250	**0.750**	**no**	0.116	+274
DS04 s44	0.000	0.000	yes	1.000	−13

Nine of the eleven ties are double zeros. On nine of the twelve jobs neither score raises a single pre-onset false alarm, so on three quarters of the panel the false-alarm comparison has no resolution: it cannot distinguish the two scores because there is nothing to distinguish. On the two ties where either score fires, the pre-onset alarm sets are identical (Jaccard 1.000): the same windows, not merely the same count. The honest statement is therefore narrower than "Mahalanobis matches $\mathcal{U}_\perp$"—the false-alarm comparison is floor-limited on nine of twelve jobs, and where it does have resolution among the ties the two scores flag exactly the same windows. The panel as a whole is not empty of information: it contains nine double zeros, two identical nonzero responses, and one job on which Mahalanobis raises three times the false-alarm rate. It also does not weaken the health-coordinate comparison, which pools a false-alarm rate of 0.292 and therefore has resolution throughout; the floor limit is specific to Mahalanobis.

Equal false-alarm rates do not imply equal timing. Across the ties, Δlead spans $[-27, +136]$ cycles with seven of eleven positive, a two-sided sign test at $p \approx 0.55$. We therefore claim no direction: the scalar false-alarm rate conceals per-job timing differences of up to a hundred cycles in both directions, which is a limitation of the summary statistic rather than a result about either score.

The one non-tie, and what we do not conclude from it. On DS04 seed 43 the Mahalanobis false-alarm rate triples (0.750 against 0.250), its first alarm sits 274 cycles further from failure, and the pre-onset alarm sets diverge (Jaccard 0.116). Two units raise pre-onset alarms under Mahalanobis and none under $\mathcal{U}_\perp$. A single-Gaussian model of the latent plausibly strains under the operating-condition heterogeneity DS04 introduces, but the other two DS04 seeds in this panel are double-zero ties, so this is one job's behaviour and not a demonstrated dataset effect. We report it as an observation and claim no mechanism.

Q Near-Failure Alarm: Cross-Subset Split

Table 29 repeats the near-failure protocol of Table 4 with the held-out-*subset* split, in which the test units come from an N-CMAPSS subset unseen in training. It is included because it exposes how the baseline coverage in Table 4 is obtained. Seven of the twelve configurations clear coverage 0.5 here, and every one of them does so at a false-alarm rate above 0.6; two reach a false-alarm rate of exactly 1.000, meaning the rolling maximum is above threshold on every pre-onset window as well as every near-failure window. Rows marked † satisfy FA ≥ 0.5 and are shaded red: their coverage carries no information, by the same argument that excludes the d=320 rows from the onset-detection ranking (Section 8.1). The two rows with genuinely low false-alarm rates, Tr.AE at d=4 (0.000) and d=2 (0.214), cover only 0.262 and 0.214 of units. No baseline representation achieves both, and our own

route is not run cross-subset for the near-failure task because the transverse geometry is refitted per subset in Stage 2.

Table 28: **Near-failure in-domain: the six baselines omitted from Table 4.** Protocol, columns and shading identical to Table 4; ranking marks are those of the full twelve-row comparison.

	DS01			DS02			DS03			DS04		
Signal	cov	FA	Lead	cov	FA	Lead	cov	FA	Lead	cov	FA	Lead
STEP $\mathcal{U}_{\text{geo}}$ (d=2)	0.33	0.08	218	0.22	0.44	320	0.00	0.33	257	0.33	0.08	228
SoftCLT-TR $\mathcal{U}_{\text{geo}}$ (d=2)	0.08	0.42	252	0.33	0.22	218	0.00	0.33	281	0.42	0.25	316
TS2Vec $\mathcal{U}_{\text{geo}}$ (d=2)	0.33	0.00	222	0.00	0.11	341	0.39	0.17	157	0.08	0.25	309
TS2Vec $\mathcal{U}_{\text{geo}}$ (d=4)	0.00	0.00	218	0.00	0.00	261	0.22	0.33	409	0.50	0.25	299
STEP $\mathcal{U}_{\text{geo}}$ (d=4)	0.08	0.25	243	0.11	0.22	228	0.00	0.33	262	0.50	0.42	319
SoftCLT-TR $\mathcal{U}_{\text{geo}}$ (d=4)	0.00	0.00	205	0.33	0.00	188	0.00	0.61	306	0.17	0.17	225

Table 29: **Near-failure alarm on the cross-subset split (baseline $\mathcal{U}_{\text{geo}}$, CNN route).** Protocol identical to Table 4. cov = coverage within the final 50 cycles, FA = pre-onset false-alarm rate, Lead = cycles from first alarm to end of life, In-dom. cov = the mean DS01–DS04 coverage of the same configuration from Table 4, for reference. Rows marked † have FA ≥ 0.5 and are shaded red; they are shown for completeness and excluded from any ranking. Sorted by cross-subset coverage.

Representation	cov ↑	FA ↓	Lead	In-dom. cov
SoftCLT-TS2Vec $\mathcal{U}_{\text{geo}}$ (d=4)†	1.000	1.000	486	0.806
TS2Vec $\mathcal{U}_{\text{geo}}$ (d=4)†	0.952	0.976	474	0.180
TS2Vec $\mathcal{U}_{\text{geo}}$ (d=2)†	0.881	1.000	497	0.201
SoftCLT-TS2Vec $\mathcal{U}_{\text{geo}}$ (d=2)†	0.857	0.881	490	0.847
BiLSTM-AE $\mathcal{U}_{\text{geo}}$ (d=4)†	0.786	0.690	377	0.403
SoftCLT-TR $\mathcal{U}_{\text{geo}}$ (d=2)†	0.714	0.619	343	0.208
SoftCLT-TR $\mathcal{U}_{\text{geo}}$ (d=4)†	0.571	0.667	373	0.125
BiLSTM-AE $\mathcal{U}_{\text{geo}}$ (d=2)	0.405	0.476	318	0.438
Tr.AE $\mathcal{U}_{\text{geo}}$ (d=4)	0.262	0.000	72	0.347
Tr.AE $\mathcal{U}_{\text{geo}}$ (d=2)	0.214	0.214	200	0.389
STEP $\mathcal{U}_{\text{geo}}$ (d=2)†	0.167	0.500	326	0.222
STEP $\mathcal{U}_{\text{geo}}$ (d=4)	0.071	0.310	284	0.173

R DIRECT $U_\perp$ VERSUS $\mathcal{U}_{\text{geo}}$: 12 MATCHED JOBS

The frozen CNN representation and geometry are held fixed; no encoder, principal curve, or RUL readout is retrained. The detector is the same rolling maximum (W=20) at the training pre-onset p_{98} threshold. As elsewhere, $\text{lead}_{\text{first}}$ is cycles from the first alarm to end of life, so a larger value means the alarm fires earlier and a smaller one that it sits tighter to failure; for a near-failure alarm the tighter value is preferable, which is why the $|h|$ control's much larger lead is a deficit rather than an advantage once its false-alarm rate is read alongside it. Here cov_H denotes coverage within the final H cycles. In these runs, $\text{cov}_{20} = \text{cov}_{50} = \text{cov}_{100}$ for every score because a covered unit remains above the rolling-max threshold through end of life (EOL); the three scores also cover the same units in every matched run.

Table 30: **Direct near-failure comparison on 12 matched jobs.** Means±sample SD over DS01–DS04, seeds 42–44. The detailed table retains cov20 and cov50 for auditability; they coincide with cov100 under this detector.

Signal	cov20	cov50	FA ↓	$\text{lead}_{\text{first}}$ ↓	first-in-50 ↑	n
$U_\perp$	0.979±0.072	0.979±0.072	0.049±0.090	114.3±71.3	0.312±0.396	12
$\mathcal{U}_{\text{geo}}$	0.979±0.072	0.979±0.072	0.083±0.107	116.5±88.2	0.361±0.396	12
$\|h\|$	0.979±0.072	0.979±0.072	0.292±0.309	242.9±94.4	0.062±0.155	12

Paired differences. For $|h| - U_\perp$, ΔFA is $+0.243 \pm 0.311$ (median $+0.208$; sign $7/5/0$) and $\Delta\text{lead}_{\text{first}}$ is $+128.7 \pm 79.3$ cycles (median $+123.5$; sign $12/0/0$), i.e. the control alarms earlier on every job while raising six times the false alarms. For $\mathcal{U}_{\text{geo}} - U_\perp$, ΔFA is $+0.035 \pm 0.083$ (median 0.000; sign $2/10/0$) and $\Delta\text{lead}_{\text{first}}$ is $+2.2 \pm 32.4$ cycles (median -4.0; sign $3/1/8$). These summaries are descriptive; pooled mean$\pm$SD differences are not treated as significance tests.

S LIFECYCLE ROLES OF h AND $\mathcal{U}_\perp$

Figure 3 contrasts what the onset and near-failure alarms measure.

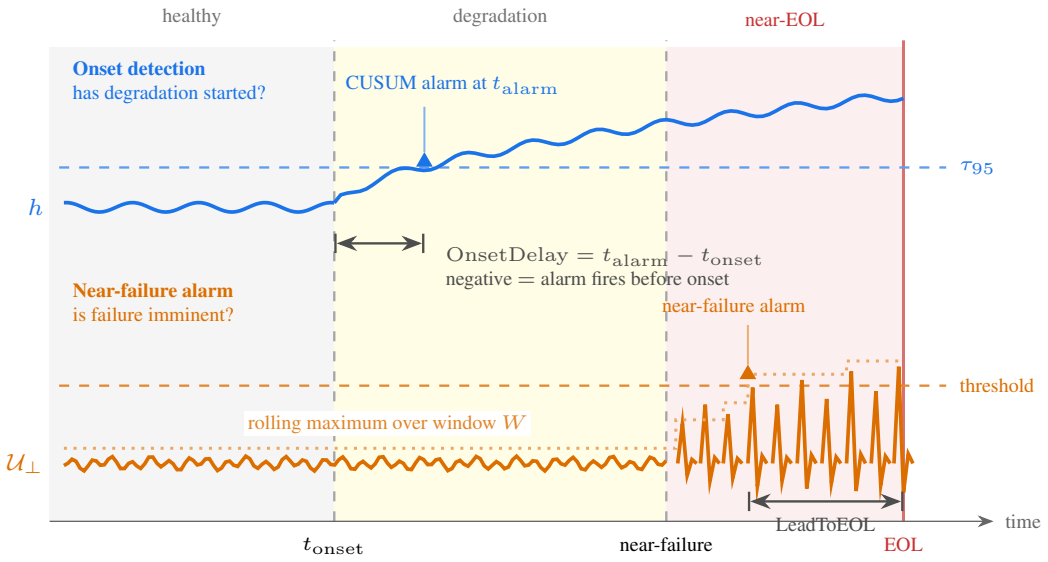

Figure 3: **What the two operational tasks measure, and why they are different questions.** Both alarms operate on a frozen representation with thresholds taken from the training distribution, but they read different signals over different parts of the lifecycle. *Top (onset):* the health signal h leaves its healthy band once degradation begins; thresholding it at τ_{95} and integrating the exceedances—with a Bernoulli CUSUM (Page, 1954) or a rolling spike rate—produces an alarm at t_{alarm}, and $\text{OnsetDelay} = t_{\text{alarm}} - t_{\text{onset}}$ is negative when the alarm precedes the labelled onset. *Bottom (near-failure):* the PCNG transverse rarity $\mathcal{U}_\perp$ stays quiet through most of the degradation phase and becomes bursty close to end of life; a rolling maximum over a window W converts those bursts into a stable alarm; LeadToEOL records how far ahead of end of life that alarm first appears, and is read as a phase marker rather than as a score. The two coordinates therefore answer complementary questions—has degradation started, and is failure imminent—rather than competing on the same one. Axes are schematic; measured values are summarized in Tables 31 and 4.

Table 31: **Onset-detection headline summary.** AUROC is reported for Page CUSUM and the causal rolling-rate detector; spike fold is the post/pre spike-rate ratio. Rows with spike fold < 2 are degenerate. Full detector/latent sweep: Appendix O.

Signal	CUSUM AUROC	Rolling AUROC	Spike fold
MonoCNN d=4, $\lvert h\rvert$	0.964	0.976	14.0
MonoCNN d=4, $\lvert s\rvert$	0.965	0.974	13.9
MonoCNN d=4, $\lvert\mathcal{U}_{\text{geo}}\rvert$	0.844	0.857	13.4
TS2Vec d=320	0.935	0.712	$1.2^\dagger$
SoftCLT-TS2Vec d=320	0.981	0.642	$1.1^\dagger$

Figure 3 is the main-paper schematic for the two operational tracks; the appendix uses the same interpretation.

T Do Different Coordinates Answer Different Questions?

Section 5.3 establishes that the transverse coordinate measures what its name claims; Section 5 establishes no statistical separability. Neither shows that *which* coordinate is extreme tells an operator *which* question to ask. We tested that directly, pre-registering the statistic, the null and the reading for every outcome before running.

Design. All three components and the Stage-1 scalar $|h|$ are scored on both detection tasks over the same 12 matched CNN jobs, with the detectors and thresholds of Section 8 held fixed and nothing retuned per score. For onset we rank by AUROC under both detectors after excluding scores that fail the spike-fold degeneracy test of Section 8.1; for the near-failure alarm we rank by false-alarm rate at matched coverage. Per job we record which score wins each task. We report win/tie/loss counts and exact sign tests over the twelve jobs for completeness, but the twelve are three seeds of each of four datasets rather than twelve independent experiments, so the p-values below overstate the evidence and we read only the win/loss pattern and the dataset-level aggregates as findings.

The pre-registered test returns a null. The registered question was whether a job's onset winner predicts its near-failure winner once the marginal preferences of each task are held fixed. The null permutes the near-failure winner across jobs, preserving both marginals and destroying only the pairing. Observed disagreement is $11/12 = 0.917$ against a permutation null of 0.904 (1000 permutations, $P(\text{null} \geq \text{obs}) = 0.68$), and the closed-form independent-pair rate is 0.903. The pairing carries no information beyond the marginals, and we report that as the pre-registered result.

A post-hoc observation, not pre-registered. The marginals themselves are nearly disjoint. $|h|$ wins onset on 10 of 12 jobs and wins the near-failure alarm outright on none; the discordant pairs run 10 to 0, an exact McNemar test at $p \approx 0.002$. The score that answers "has degradation started" is systematically not the score that answers "is failure imminent". This supports the premise that one scalar cannot serve both questions, which is the motivation in Section 1. It does not show that the PCNG decomposition is what supplies the division of labour, because the onset winner is the Stage-1 scalar rather than one of the three components. We flag it as post-hoc because our registered null held the marginals fixed by construction and therefore could not have detected it; a reader should weight it accordingly.

Table 32: **Operational usability of the three components**, 12 matched jobs, both tasks. Degeneracy is the spike-fold test of Section 8.1; a score failing it attains an onset AUROC while discriminating nothing.

Component	Onset	Near-failure alarm
$\mathcal{U}_\perp$	AUROC 0.57–0.91, never degenerate	FA 0.000–0.250; wins or ties $11/12$
$\mathcal{U}_\parallel$	degenerate on $9/12$; AUROC 0.59–0.65 elsewhere	FA $= 1.000$ on $9/12$
$\mathcal{U}_{\dot{s}}$	degenerate on $12/12$	coverage 0 on $8/12$

Only one component thresholds into a usable detector. This is the operationally important result of the experiment and it is negative. $\mathcal{U}_\parallel$ fires almost continuously and $\mathcal{U}_{\dot{s}}$ misses the near-failure window outright on two thirds of jobs; on onset, $\mathcal{U}_\parallel$ is degenerate on nine of twelve jobs and $\mathcal{U}_{\dot{s}}$ on all twelve. The three-coordinate frame supports exactly one detection decision, through $\mathcal{U}_\perp$. Nothing in the paper claims otherwise—the abstract commits only $\mathcal{U}_\perp$ to the near-failure alarm and Section 5.3 already declines a rate interpretation—but the scope is narrower than a reader might infer from a three-coordinate measurement frame, so we state it: the other two coordinates are measurements, not alarms.

One case is worth naming because it validates the degeneracy test on a score it was not designed for. On DS02 seed 44, $\mathcal{U}_{\dot{s}}$ reaches an onset AUROC of 0.999 under the CUSUM detector and collapses to 0.549 under the rolling-rate detector with a spike fold of 0.89. That is the failure mode Section 8.1 documents, caught here on a PCNG component.

U Does Transverse Rarity Flag Windows Where the Predictor Is Worse?

Section 8.2 shows that $\mathcal{U}_\perp$ supports a near-failure alarm. It does not show that a window scored as novel is one on which the frozen predictor is less accurate. Those are different claims, and the second is the one a reader is entitled to ask about given a term like *predictive novelty*. We test it directly and report a mixed result.

U.1 Why the test must be conditioned on progression

Both $\mathcal{U}_\perp$ and absolute RUL error rise with life stage, so an unconditioned comparison of error between high- and low-novelty windows measures ageing. We therefore stratify the test windows into five quantile bins of s and perform selective prediction *within* each stratum: for a score ϕ, windows are ranked ascending in ϕ, the least novel fraction c is retained, and the risk at coverage c is the mean of $|\hat{y} - y|$ over retained windows. The risk–coverage curve and the selective-prediction framing are due to El-Yaniv & Wiener (2010); Geifman & El-Yaniv (2017). Stratum curves are combined at each c weighted by stratum size, and the summary statistic is the area under the resulting risk–coverage curve (AURC; lower is better). We also report excess AURC (E-AURC), the excess over an oracle that ranks by the true error, which normalises across jobs with different base error; both summaries are those of Geifman et al. (2019). The pitfalls of this evaluation—above all that it inherits every confound of the underlying task—are documented by Ding et al. (2020) and Traub et al. (2024), and the stratification below exists because of them. Twelve matched CNN jobs (DS01–DS04, seeds 42–44) give a paired design. We use an exact sign test on the non-tied jobs, with the same caveat as Appendix P: three seeds of one dataset are not three independent experiments, so these p-values are descriptive summaries of a win/loss pattern rather than dataset-level inference.

Table 33: **Progression-conditioned selective prediction.** Pooled over 12 matched jobs. Δ is the paired difference $\mathrm{AURC}(\phi) - \mathrm{AURC}(\mathcal{U}_\perp)$, so a positive value means $\mathcal{U}_\perp$ selects better; p is a two-sided exact sign test on non-tied jobs, descriptive only. The row that matters is *random*: after stratification it, not s, is the meaningful null.

Score ϕ	AURC ↓	E-AURC	Δ vs $\mathcal{U}_\perp$	$+/0/-$	p
$\mathcal{U}_\perp$	$\mathbf{9.85 \pm 6.05}$	$\mathbf{4.46 \pm 2.49}$	—	—	—
KDE NLL	9.88 ± 6.03	4.49 ± 2.48	$+0.02 \pm 0.24$	$6/0/6$	1.00
kNN distance	9.89 ± 5.97	4.50 ± 2.41	$+0.04 \pm 0.16$	$7/0/5$	0.77
$\lvert h \rvert$ control	10.03 ± 6.00	4.64 ± 2.46	$+0.18 \pm 0.38$	$10/0/2$	**0.039**
Mahalanobis	10.05 ± 5.94	4.66 ± 2.39	$+0.20 \pm 0.40$	$9/0/3$	0.15
random	10.29 ± 6.56	4.90 ± 3.01	$+0.44 \pm 0.77$	$8/0/4$	0.39
s alone	10.63 ± 6.35	5.24 ± 2.77	$+0.78 \pm 0.45$	$12/0/0$	0.0005

U.2 What we find, and what we decline to claim

$\mathcal{U}_\perp$ has the lowest pooled AURC of every score tested and improves on the Stage-1 health coordinate on 10 of 12 matched jobs. That is a genuine if modest result: the geometric score selects better than the scalar our own Stage 1 produces.

It is not separable from random selection. The paired margin is $+0.44 \pm 0.77$ with sign $8/0/4$ and $p = 0.39$; a 95% interval on the margin includes zero. Nor is it separable from Mahalanobis distance ($9/0/3$, $p = 0.15$), kNN distance ($7/0/5$) or kernel density ($6/0/6$). We therefore do *not* claim that transverse rarity identifies windows on which the predictor is less trustworthy. At 12 matched jobs this experiment can detect an effect of roughly the size we see only if it is consistent, and it is not.

Why we report the comparison against s at all. The $12/0/0$ sweep against selection on s looks like the strongest row in the table and is the weakest. After stratification, s varies little within a bin, and what variation remains is anti-correlated with absolute RUL error, because a window lower in a bin has more remaining life and therefore larger absolute error. Selection on s is consequently *worse than random* (10.63 against 10.29), and beating it establishes nothing about $\mathcal{U}_\perp$. We report the row to document that progression is not a usable selection signal once life stage is controlled, and because omitting a baseline that behaves strangely would be the kind of choice this paper exists to avoid.

The unstratified comparison, for contrast. Without stratification the margin against s inflates to -2.21 ± 1.18 ($11/12$ jobs), roughly three times the stratified value. That gap is the life-stage confound, and it is why we report only the stratified analysis as the result.

Why we do not call this predictive novelty. The coordinates measure novelty in the dimensions a predictor reads. This experiment asked whether that novelty additionally tracks prediction error, and the answer at this sample size is that it does so no better than chance, and no worse than classical latent novelty scores, which also fail to separate from chance here. A larger matched design, or a task where errors are less dominated by life stage, would be needed to settle it.

U.3 Does it track error anywhere? A stratified null

The pooled result above leaves one rescue open: transverse rarity might track error late in life, where degradation is pronounced, and be uninformative earlier. We tested that by splitting the five progression strata into three groups and repeating the paired comparison within each, pre-registering the reading before running. The pooled stratified AURC for $\mathcal{U}_\perp$ reproduces the value in Table 33 to four decimals (9.8539), which confirms the decomposition is of the same analysis. Table 34 reports the per-stratum comparisons.

Table 34: **Selective prediction by progression stratum**, 12 matched jobs. A positive Δ means $\mathcal{U}_\perp$ has the lower AURC and therefore selects better. p is an exact sign test on non-tied jobs. Early is bins 1–2, mid bin 3, late bins 4–5.

	vs. $\lvert h\rvert$ control		vs. random	
Stratum	ΔAURC	sign, p	ΔAURC	sign, p
Early	$+0.058$	$5/7$, 0.774	$+0.692$	$9/3$, 0.146
Mid	$+0.329$	$9/3$, 0.146	$+0.451$	$6/6$, 1.00
Late	$+0.216$	$\mathbf{10/2}$, $\mathbf{0.039}$	$+0.182$	$7/5$, 0.774

No stratum clears both controls. $\mathcal{U}_\perp$ separates from the health coordinate only late in life ($10/2$, $p = 0.039$) and from random selection nowhere—$9/3$, $6/6$ and $7/5$, with sign-test p of 0.146, 1.00 and 0.774. As above, these are 12 dependent jobs per stratum, so we read the counts and not a threshold.

The late-stratum sign count coincides with the pooled one in Table 33 but is not the same effect. Recombining the strata at their bin weights reproduces the pooled margins ($+0.176$ against $|h|$, $+0.440$ against random, against $+0.18$ and $+0.44$ reported), and the two jobs on which $|h|$ wins pooled—DS02 seed 44 and DS04 seed 44—are not the two on which it wins in the late stratum, which are DS01 seeds 42 and 44. The health-coordinate advantage is present in mean at every stage ($+0.058$, $+0.329$, $+0.216$), is largest in the middle and shows its most consistent sign pattern late, and the pooled comparison accumulates that advantage across strata rather than drawing it from any one of them. The stratification therefore does not localise the effect; it shows it to be small and diffuse. The rescue is therefore closed in the direction it was offered: there is no life stage at which transverse rarity becomes a usable predictor of where the frozen model errs. We state this as a limit of this experiment at 12 jobs per stratum rather than as a general impossibility, but the claim the paper declines to make is now declined with evidence at every stage of life rather than only on average.

V Scalar Ambiguity at Matched Mahalanobis Novelty

A scalar novelty score can contain some information about mechanism, so a fair attribution test must not assume that equal scalar novelty is mechanism-blind. We therefore ask a narrower question on the frozen MonoCNN packs (DS01–DS03, seeds 42–44): among windows with similar mid–high Mahalanobis novelty, do $\mathcal{U}_\perp$-dominant and $\mathcal{U}_\parallel$-dominant windows occupy different situations? No encoder, curve, or predictor is retrained. Write M for the Mahalanobis novelty score on the frozen embedding. The locked SUPPORT criterion requires at least two thirds of jobs to show both higher mean life fraction and higher late-life rate for the $\mathcal{U}_\parallel$-dominant group, together with mean Cliff's $\delta > 0.1$.

The locked band is $M \in [q_{0.6}, q_{0.95}]$ with dominance $|\mathcal{U}_\perp - \mathcal{U}_\parallel| \geq 0.5$ and late life defined as life fraction ≥ 0.8. The primary reading applies that SUPPORT criterion. A cleaner follow-up greedily matches each $\mathcal{U}_\perp$-dominant window one-to-one to the nearest $\mathcal{U}_\parallel$-dominant window in Mahalanobis score, without replacement, under $|\Delta M| \leq 0.05 \times \text{median}(|M|_{\text{band}})$.

Table 35: **Scalar-ambiguity result: PARTIAL under both readings.** The nearest-M analysis removes the visible Mahalanobis imbalance in the band contrast but does not make the locked primary criterion pass.

Analysis	late-rate dir.	mean-life dir.	mean Cliff δ	frac. pairs $\parallel$ later
Within-band groups	$9/9$	$5/9$	$+0.401$	—
Nearest-M pairs	$9/9$	$5/9$	$+0.449$	0.76

The matched analysis is the cleaner test of "same scalar novelty, different situation": mean $|\Delta M|$ is between 0.0042 and 0.0103 across jobs and the mean signed ΔM is within 0.0034 of zero in every job. Despite that balance, the $\mathcal{U}_\parallel$-dominant member is later in the majority of matched pairs in all $9/9$ jobs, and the late-life rate is higher in all $9/9$; mean Cliff's $\delta = +0.449$. Mean life fraction, however, increases in only $5/9$ jobs, short of the locked $\geq 2/3$ criterion. We label the experiment **PARTIAL**, not SUPPORT.

The locked SUPPORT criterion is therefore not met. We report the experiment as PARTIAL and do not use it as evidence for scalar insufficiency or detector superiority. An earlier equivalence experiment that directly matched mechanisms on Mahalanobis was withdrawn because Mahalanobis itself partly predicted mechanism; that negative result is why the present analysis asks only about situational labels within a controlled scalar-novelty band.

W SCHEMATIC BACKBONE GEOMETRY

Transverse geometry is backbone-dependent: the two encoders we contribute to the pool place their transverse residual differently, which is why the near-failure alarm study is reported on the convolutional route only. The following schematic illustrates the two observed regimes; plotted points are illustrative rather than empirical embeddings.

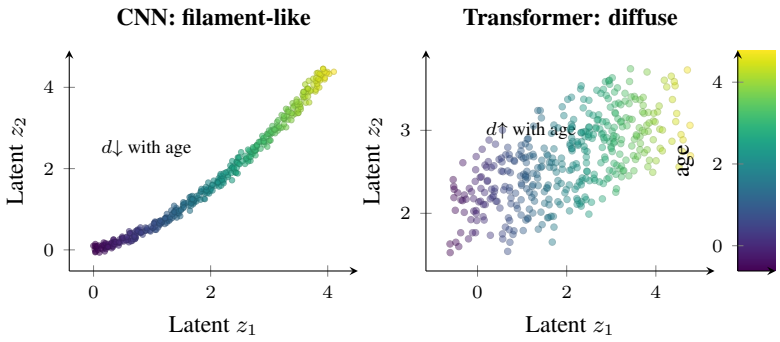

Figure 4: **Schematic illustration of two backbone-dependent transverse regimes.** Plotted points are illustrative, not empirical embeddings. The CNN panel shows a filament-like organisation with d decreasing in age; the Transformer panel shows a diffuse cloud with d increasing in age. Exact PCA or health-correlation annotations are omitted because this figure is not a measured embedding plot.