BEYOND COVARIANCE: HOW SHALLOW AUTOENCODERS LEARN NORMAL RELATIONS FOR MULTIVARIATE TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A simple shallow autoencoder (ShallowAE) is highly competitive on the TSB-AD-M benchmark for multivariate time series anomaly detection, yet what it learns from normal data and how this information shapes anomaly scores remain unclear. We characterize the statistical sufficiency boundary of squared-error reconstruction objectives, proving that the mean and covariance determine the entire risk function for affine reconstruction but are not universally sufficient for nonlinear reconstruction objectives. Moment-matched interventions on real time series show that training information beyond second-order statistics can change ShallowAE's learning and detection behavior. Controls within a fixed PCA subspace further show that the network can assign different scores to different combinations of variables even when all inputs lie in the same subspace and have zero PCA residual. An analytically tractable relation-switching experiment reveals that, between normal environments with identical coordinate-wise marginals and covariance, reversing the conditional relationship changes both the network's low-residual regions and its scoring preferences on fixed inputs. These findings identify the learning of normal relationships beyond covariance as a mechanism through which ShallowAE can reshape reconstruction geometry and score inputs differently despite identical PCA reconstruction scores.

1 INTRODUCTION

Multivariate time-series anomaly detection often has abundant normal records but few representative anomalies (Pang et al., 2021). Reconstruction-based detectors therefore learn to restore normal windows and use reconstruction error to flag unusual inputs (Audibert et al., 2020). Their success requires a gap between normal and anomalous reconstruction errors, yet anomalies can also be reconstructed accurately (Gong et al., 2019; Yoon et al., 2021). Low training error characterizes fit to observed normal data, whereas detection depends on the reconstruction rule's behavior on other inputs. Explaining the error gap therefore requires identifying which normal training information shapes the learned rule.

Simple baselines make this question concrete. Comparative studies find no consistent advantage of complex architectures over conventional methods (Audibert et al., 2022; Liu & Paparrizos, 2024). In Figure 1, ShallowAE and global window PCA (GWPCA) rank first and second by VUS-PR (Boniol et al., 2025) over 180 sequences from the TSB-AD-M benchmark (Liu & Paparrizos, 2024). Both reconstruct standardized windows containing temporal and channel information and score reconstruction residuals. GWPCA determines a single principal subspace from the global mean and covariance; ShallowAE can use an input-dependent reconstruction map. Their common representation and scoring principle provide a reference for asking which information the nonlinear model uses beyond the statistics that determine linear reconstruction. The benchmark motivates this comparison without identifying the source of its performance.

Nonlinear autoencoders can represent curved variable relations (Kramer, 1991), but this capacity leaves open which distributional differences affect a trained detector and how they change anomaly rankings. We ask which normal joint information changes reconstruction learning when global second-order statistics are held fixed, and whether the resulting score variation extends beyond dis-

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

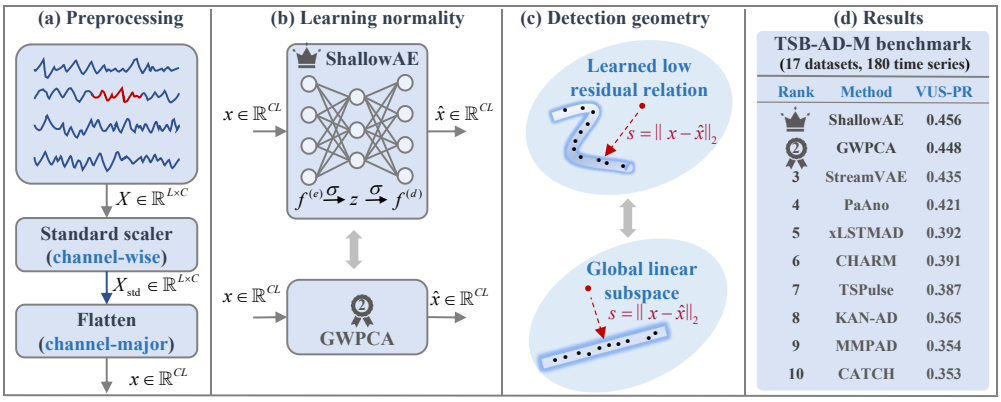


Figure 1: Window reconstruction and motivating results. (a) Channel-wise standardization gives X_{std}, followed by channel-major flattening. (b) ShallowAE and GWPCA learn nonlinear and global linear reconstruction maps, respectively. (c) Both use $s = \|x-\hat{x}\|_2$, but induce different low-residual geometries. (d) ShallowAE and GWPCA rank first and second by sequence-averaged VUS-PR on TSB-AD-M (17 datasets, 180 series). Full rankings and aggregation details appear in Appendix B.

tance to a PCA subspace. We combine moment-matched training interventions with fixed-model queries at equal PCA residual, separating sensitivity to training information from sensitivity to subspace distance. An analytical construction then changes a specified conditional relation while preserving coordinate-wise marginals and covariance, allowing its effect on reconstruction and scoring to be tested directly.

The evidence identifies selectivity for combinations within a subspace as an additional degree of freedom in scoring: ShallowAE assigns different scores to inputs with identical PCA residuals. In the analytical environments, reversing the normal conditional relation produces bidirectional changes in learned low-residual regions and scoring preferences on fixed queries, directly linking a specified change in training information to reconstruction geometry and anomaly separation. This selectivity has detection value when the learned combinations remain valid for future normal variation and anomalies deviate from them. Real-data interventions show reproducible responses across random transforms and variation in this alignment across settings.

Our contributions span practical baselines, a statistical characterization, and controlled mechanism evidence. ShallowAE and GWPCA provide simple window-reconstruction baselines for multivariate time-series detection. We characterize when global first and second moments universally determine an expected squared reconstruction score: precisely when that score is globally quadratic on the stated domain. This yields affine-objective invariance, while activation-conditioned moments describe the finer information dependence of ReLU reconstruction. Controlled experiments connect that dependence to within-subspace score selectivity and relation-dependent scoring. Conditional-mean theory specifies the target of an optimal decoder at a fixed representation, and an approximation bound states when relation violations suffice for score separation. These results connect training statistics to conditions under which departures from learned normal relations are score-separable.

2 Related Work

Relationship modeling and reconstruction-based detection. Graph-based and frequency dependent models use variable dependencies for anomaly detection (Deng & Hooi, 2021; Wu et al., 2025); MemAE and NAE constrain anomaly reconstruction through memory mechanisms and energy modeling, respectively (Gong et al., 2019; Yoon et al., 2021). These works emphasize the importance of modeling relationships and constraining anomaly reconstruction. We study the connection between the two: whether the joint organization of normal variables still changes reconstruction learning when global second-order statistics are held fixed, and how such changes affect anomaly scores.

Theories of reconstruction learning and anomaly separation. Optimal reconstruction in linear autoencoders is linked to PCA (Baldi & Hornik, 1989), whereas nonlinear autoassociative networks can represent curved variable relations (Kramer, 1991). Nonlinear representation capacity leaves open which training information determines anomaly scores. Alain and Bengio (2014) connect optimal reconstruction maps to the gradient of the log data density under a specific reconstruction-contractive criterion in the vanishing-regularization limit; Chen and Liu (2026) explain how learning dynamics produce reconstruction gaps under an orthogonal image-patch feature model. We approach the problem through statistical sufficiency, characterizing when global second-order statistics suffice to determine squared reconstruction risk and using controlled experiments to examine how conditional relations beyond these statistics change anomaly separation.

Statistical controls. Surrogate-data tests preserve selected statistics to test whether time series contain structure unexplained by the corresponding reference model (Theiler et al., 1992; Schreiber & Schmitz, 2000). We adopt this control principle to intervene on normal training distributions and examine the responses of reconstruction learning and anomaly scores. Because moment matching can change several joint statistics simultaneously, we combine fixed-subspace controls with analytical relation constructions to delimit the interpretation of specific mechanisms.

3 From Training Statistics to Anomaly Separation

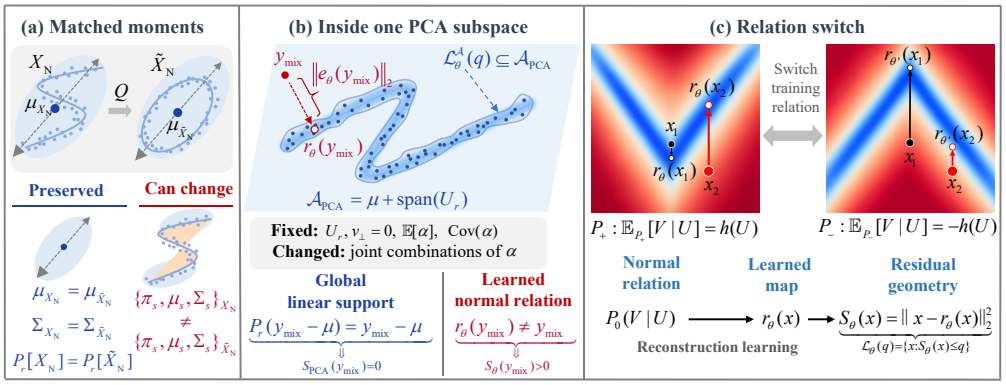


Figure 2: Identifying normal relations beyond covariance. (a) Orthogonal mixing of X_{N} preserves the training mean, covariance, and PCA projector (when unique), while activation-conditioned statistics may change. $P_r[\cdot]$ denotes the projector fitted to the indicated data. (b) The query $y_{\mathrm{mix}} = \mu + U_r \alpha_{\mathrm{mix}}$ lies in the PCA support, with projector $P_r = U_r U_r^\top$; $\mathcal{L}_\theta^{\mathcal{A}}(q)$ is the low-residual set restricted to this support. Filled/open points denote inputs/reconstructions; connecting lines represent residuals. (c) Maps $r_\theta, r_{\theta'}$ learned under P_+, P_- are compared on fixed queries x_1, x_2. The chain $P_0(V \mid U) \to r_\theta(x) \to S_\theta(x)$ summarizes the investigated link from normal training relations to reconstruction and anomaly scores, with P_0 denoting the normal training environment.

3.1 Window Reconstruction and a Common Reference

As Figure 1(a) shows, a window $X_t \in \mathbb{R}^{L \times C}$ of length L with C channels is standardized using channel means and standard deviations estimated from normal training data to give $X_{t,\mathrm{std}}$ (Figure 1 omits t), then flattened in channel-major order to $x_t \in \mathbb{R}^D$, where $D = LC$. Compared with a pointwise autoencoder that receives only one time step, this representation allows fully connected layers to jointly model combinations across times and channels within a fixed window. In Figure 1(b), ShallowAE has the symmetric architecture $D \to H \to d \to H \to D$, where H is the hidden width and $d < D$ is the bottleneck dimension. The encoder $f_{\theta_e}^{(e)} : \mathbb{R}^D \to \mathbb{R}^d$ and decoder $f_{\theta_d}^{(d)} : \mathbb{R}^d \to \mathbb{R}^D$ each have one hidden layer with activation σ; the bottleneck and output layers have no activation. The bottleneck representation is $z = f_{\theta_e}^{(e)}(x)$, and $\theta = (\theta_e, \theta_d)$ collects the model parameters.

ShallowAE trains only on normal windows. Its reconstruction map, residual, and the squared score used in our analysis are

$$r_\theta = f_{\theta_d}^{(d)} \circ f_{\theta_e}^{(e)}, \qquad e_\theta(x) = x - r_\theta(x), \qquad S_\theta(x) = \|e_\theta(x)\|_2^2. \tag{1}$$

Window residual scores are aligned to the time axis at window centers. Training minimizes the empirical mean of S_θ/D. GWPCA uses the same standardized windows and residual scoring, reconstructing by projection onto the principal subspace of the training mean and covariance. Lagged-window PCA has precedents in statistical monitoring (Camacho et al., 2016). This reference fixes the input representation while exposing the information dependence of linear reconstruction.

3.2 How information beyond second-order statistics enters learning

We first establish which statistics suffice to keep the reference model's entire objective unchanged. Let $\Omega \subseteq \mathbb{R}^D$ contain a nonempty open set, and let P, P' be distributions on Ω with finite second moments. Write $P \equiv_2 P'$ when their means and raw second moments agree, equivalently when their means and full covariances agree.

Theorem 1 (Characterization of second-moment sufficiency)**.** *For a fixed reconstructor, if S_θ is a polynomial of degree at most two on Ω, then $\mathbb{E}_P S_\theta = \mathbb{E}_{P'} S_\theta$ for every $P \equiv_2 P'$. Conversely, if the score is not in this polynomial space, there exist two moment-matched finite-support distributions with different expected scores.* Proof in Appendix A.1.

For affine reconstructors, including deep linear AEs, matching the mean and covariance preserves the entire squared risk and, wherever differentiable, population and empirical full-batch gradients; individual mini-batch gradients need not agree (Appendix A.2). The converse establishes the existence of distinguishable moment-matched distributions for nonquadratic scores, not a detection benefit for every intervention.

We collect n normal windows into $X_{\mathrm{N}} \in \mathbb{R}^{n \times D}$ and apply a transform along the sample axis:

$$\widetilde{X}_{\mathrm{N}} = Q X_{\mathrm{N}}, \qquad Q^\top Q = I_n, \qquad Q 1_n = 1_n. \tag{2}$$

The subscript N identifies normal training data; $Q \in \mathbb{R}^{n \times n}$ is a transform matrix and $1_n \in \mathbb{R}^n$ is the all-ones vector. This transform exactly preserves the empirical mean and full covariance, giving $\mu_{X_{\mathrm{N}}} = \mu_{\widetilde{X}_{\mathrm{N}}}$ and $\Sigma_{X_{\mathrm{N}}} = \Sigma_{\widetilde{X}_{\mathrm{N}}}$ in Figure 2(a). When the principal subspace at the truncation rank is unique, the exact PCA projector $P_r = U_r U_r^\top$ is also unchanged; the columns of U_r are the leading r orthonormal principal directions. Experiments additionally verify projector and score invariance. The transform allows marginals and multiple joint relations to change together, so this control identifies the aggregate role of information beyond second-order statistics. The later analytical construction isolates a specific relation.

The piecewise-affine structure of ReLU networks (Balestriero & Baraniuk, 2018) further explains how such changes enter the learning objective. At fixed parameters, $r_\theta(x) = A_s x + b_s$ within activation region $\mathcal{R}_s(\theta)$. For the normal random vector X, define regional raw moments

$$M_s^{(0)} = P(X \in \mathcal{R}_s), \qquad M_s^{(1)} = \mathbb{E}[X 1_{\mathcal{R}_s}(X)], \qquad M_s^{(2)} = \mathbb{E}[X X^\top 1_{\mathcal{R}_s}(X)], \tag{3}$$

where $1_{\mathcal{R}_s}(X)$ is the region indicator. In Figure 2(a), π_s, μ_s, Σ_s denote regional probability, conditional mean, and conditional covariance. For $\pi_s > 0$, the three raw moments equal π_s, $\pi_s \mu_s$, and $\pi_s(\Sigma_s + \mu_s \mu_s^\top)$, respectively; zero-probability regions have zero raw moments.

Theorem 2 (Sufficiency of activation-conditioned moments)**.** *Let $C_s = I - A_s$. The squared risk of a fixed ReLU reconstructor is*

$$\mathbb{E} S_\theta(X) = \sum_s \left[\mathrm{tr}(C_s M_s^{(2)} C_s^\top) - 2 b_s^\top C_s M_s^{(1)} + \|b_s\|_2^2 M_s^{(0)} \right]. \tag{4}$$

For a finite training set with all hidden-unit preactivations nonzero, there is a parameter neighborhood with unchanged activation patterns in which the empirical risk and full-batch gradient are determined jointly by empirical regional moments and model parameters. Proof in Appendix A.3.

Global moments aggregate regional moments without retaining their allocation across input space. We compare risks and gradients at shared weights, then train paired models evaluated on common

test data. For $\mathcal{D} = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ and $\lambda \in [0, 1]$, define

$$L_\lambda(\theta) = (1-\lambda)L(\theta; X_{\mathrm{N}}) + \lambda L(\theta; \widetilde{X}_{\mathrm{N}}), \qquad L(\theta; \mathcal{D}) := \frac{1}{nD}\sum_{i=1}^{n} \|x_i - r_\theta(x_i)\|_2^2. \tag{5}$$

At each update, equal-sized original and moment-matched mini-batches use the same weighting; $\lambda = 0$ and 1 recover original-only and matched-only training. Paired runs share initialization, batch order, and update count, while continuation from shared parents tests whether later training rewrites the learned behavior.

3.3 Locating score selectivity within a fixed subspace

To locate the input differences used for scoring, we fix the trained model and vary queries at equal PCA residual. Let $U_r \in \mathbb{R}^{D\times r}$ be the original normal windows' PCA basis, μ their mean, and $\mathcal{A}_{\mathrm{PCA}} = \mu + \mathrm{span}(U_r)$. For coefficients $\alpha \in \mathbb{R}^r$ and orthogonal residual $v_\perp \in \mathbb{R}^D$,

$$y = \mu + U_r\alpha + v_\perp, \qquad U_r^\top U_r = I_r, \quad U_r^\top v_\perp = 0, \qquad S_{\mathrm{PCA}}(y) = \|v_\perp\|_2^2. \tag{6}$$

The PCA score is independent of the combination of coefficients α, as proved in Corollary 1.2 in Appendix A.2. We fix $v_\perp$ and change joint combinations while preserving coefficient means and covariances. Queries lie entirely within the same subspace when $v_\perp = 0$; a fixed nonzero $v_\perp$ controls a common off-subspace deviation. The additional marginal-constrained construction exactly preserves empirical coefficient marginals and matches covariance and correlation moments approximately, so we interpret it separately from the first two exact controls. Controlled queries test score sensitivity to internal combinations, complemented by an observational check on real anomaly events with approximately equal PCA scores; protocols are in Appendix D.

Given a normal-calibration threshold q, define the low-residual set and its PCA-support restriction:

$$\mathcal{L}_\theta(q) = \{x : S_\theta(x) \le q\}, \qquad \mathcal{L}_\theta^{\mathcal{A}}(q) = \mathcal{L}_\theta(q) \cap \mathcal{A}_{\mathrm{PCA}}. \tag{7}$$

PCA gives zero residual throughout $\mathcal{A}_{\mathrm{PCA}}$, whereas $\mathcal{L}_\theta^{\mathcal{A}}(q)$ may cover only part of it (Figure 2(b)). Unequal network scores at equal PCA scores reveal within-subspace selectivity; the next construction identifies a relation sufficient to produce it.

3.4 How conditional relations shape reconstruction and anomaly separation

Regional moments specify how the training distribution enters a fixed network's risk. Conditional-mean theory addresses a different question: which reconstruction is optimal for a given representation? This distinction separates the source of learning signals from the decoder's prediction target. We use the classical conditional-expectation result to define that target.

Lemma 3 (Optimal decoder for a fixed representation)**.** *Suppose $\mathbb{E}\|X\|_2^2 < \infty$. Fix the encoder and let $Z = f_{\theta_e}^{(e)}(X)$. Among measurable decoders satisfying $\mathbb{E}\|g(Z)\|_2^2 < \infty$, a squared-risk minimizer is $g_P^*(z) = \mathbb{E}_P[X \mid Z = z]$, with equality almost everywhere under the distribution of Z.*

Proof in Appendix A.4. The lemma specifies a conditional-mean target, related to principal-curve self-consistency (Hastie & Stuetzle, 1989). Its scope is the optimal decoder for a fixed representation; the behavior of a finite network that jointly learns its encoder and decoder is tested experimentally. We construct environments with identical global statistics but opposite optimal relations for a specified representation.

As in Figure 2(c), we isolate the role of a conditional relation using two environments:

$$\begin{aligned} &U \sim \mathrm{Unif}[-1, 1], \quad h(U) = |U| - \tfrac{1}{2}, \\ &\eta \sim \mathrm{Unif}[-\delta, \delta], \quad V_\kappa = \kappa\, h(U) + \eta, \quad \kappa \in \{-1, +1\}. \end{aligned} \tag{8}$$

Here U is scalar, η is independent of U, $\delta > 0$ is the noise amplitude, and P_κ is the distribution of (U, V_κ). Analytical risks use the original two-dimensional coordinates.

Theorem 4 (Identical marginals and covariance, opposite optimal relations)**.** *The distributions P_+ and P_- have identical coordinate-wise marginals, zero means, and full covariance, but $\mathbb{E}[V_\kappa \mid U =$*

$u] = \kappa h(u)$. *By Lemma 3, within the class of relation reconstructors* $r_m(u, v) = (u, m(u))$*, the optimal relation is* $m^*_\kappa(u) = \kappa h(u)$*. Writing* $R_P(m) = \mathbb{E}_P[(V - m(U))^2]$*, we have*

$$R_{P_{-\kappa}}(m^*_\kappa) - R_{P_{-\kappa}}(m^*_{-\kappa}) = 4\mathbb{E}[h(U)^2] = \frac{1}{3}. \tag{9}$$

For the ideal relation reconstructor of environment P_κ, the residual in its own environment is only the noise η; in the other environment it additionally contains the relation shift $-2\kappa h(U)$. The gap between the two conditional relations therefore supplies extra reconstruction error, whose average risk cost is $1/3$. The common covariance, noise risk $\delta^2/3$, and proof are given in Appendix A.5; experimental coordinates and scoring conventions are in Appendix E.1.

To test whether a trained ShallowAE responds accordingly, we branch from the same parent model and continue on the retained or switched environment, matching optimizer resets and training budgets. The retained branch controls for continued optimization. On fixed queries, we measure whether switching redirects both reconstructed outputs toward the target relation and score preferences. Shared-scale fields show geometry; rankings on the common test set measure separation.

Score separation additionally depends on relation deviation, normal noise, and approximation error. The following condition compares these quantities.

Corollary 5 (Score-separation conditions from relation approximation)**.** *Let* $r_*(u, v) = (u, f(u))$*, and suppose the learned reconstructor satisfies* $\sup_{x\in\Omega} \|r_\theta(x) - r_*(x)\|_2 \le \varepsilon$ *on a query domain* Ω *containing both normal and anomalous inputs. Define the fixed-*u *deviation from the relation as* $d_f(x) = \|v - f(u)\|_2$*. Then*

$$[d_f(x) - \varepsilon]_+ \le \sqrt{S_\theta(x)} \le d_f(x) + \varepsilon, \tag{10}$$

where $[w]_+ = \max(w, 0)$*. If normal inputs satisfy* $d_f(x) \le \delta$ *and anomalous inputs satisfy* $d_f(x) \ge a$*, then* $a > \delta + 2\varepsilon$ *suffices for strict score separation. For the alternative environment in Theorem 4, on the clear region* $|h(U)| \ge b$*, a sufficient condition is* $b > \delta + \varepsilon$*; at the intersections* $h(U) = 0$*, the two conditional distributions coincide.* Proof in Appendix A.6.

Here d_f measures fixed-u deviation; a bounds anomaly deviation from below, b bounds relation amplitude, and δ bounds normal noise. All use the same coordinate scale. The approximation assumption applies uniformly over the query domain.

Under controlled approximation error, relation violations satisfying this condition are separable by reconstruction score. Path barriers and the residual-response identity provide additional geometric diagnostics (Appendix A.7).

4 Experiments

Figure 1 establishes the strong-baseline starting point; the experiments below ask how normal training information enters reconstruction learning and affects anomaly scores and detection. Except for the analytical toy, the mechanism experiments use window length 100, hidden width 64, bottleneck dimension 16, ReLU, and three training seeds. Their cross-sequence effects weight dataset families equally. The separate parameter scan uses seed 2026 and sequence-equal aggregation. Complete benchmark and parameter-sensitivity results are in Appendix B, and experimental protocols are in Appendices C–E.

4.1 Reconstruction learning beyond second-order statistics

After matching the mean and full covariance of normal windows, we compare learning objectives for original and moment-matched inputs at the same weights. In Figure 3(a), the linear AE's risk and full-batch gradient remain numerically identical, while nonlinear ShallowAE's gradient changes. This contrast realizes the statistical sufficiency boundary: the ReLU activation-conditioned moment decomposition specifies how information omitted by global moments enters learning. At the same optimization starting point, the two training distributions already provide different update signals, independently of post-training performance.

This objective difference also carries through to detection. In the original six-sequence mechanism set, original training outperforms moment-matched training on all six sequences, while exact PCA

remains unchanged. The near-zero reordering effect and the retained advantage under deterministic full-batch training attribute the observed difference to the training distribution beyond mini-batch order. Table 1 summarizes the paired effects; Appendices C.1–C.2 report the objective-sensitivity and optimization controls.

4.2 Repeatability and Partial Recovery of Detection Behavior

Figure 3(b) extends the test to 11 sequences from eight dataset families. Three independent moment-matching transforms yield the same effect sign on 11/11 sequences, establishing within-setting repeatability across transforms. Effects are positive in some settings and negative in others, so the cross-family average remains inconclusive (Appendix C.3).

In Figure 3(c), family-equal performance on the original six sequences declines as the moment-matched training fraction increases and flattens at high fractions, showing that detection behavior changes as training information is progressively altered. Figure 3(d) branches from identical parent models: both paired effects start at zero and then diverge with the continuation data. Re-exposure to the original relations recovers part of the detection gap, with a residual gap after the fixed continuation budget. The evidence therefore extends from endpoint comparisons to the learning process; Appendix C.4 reports the full trajectories.

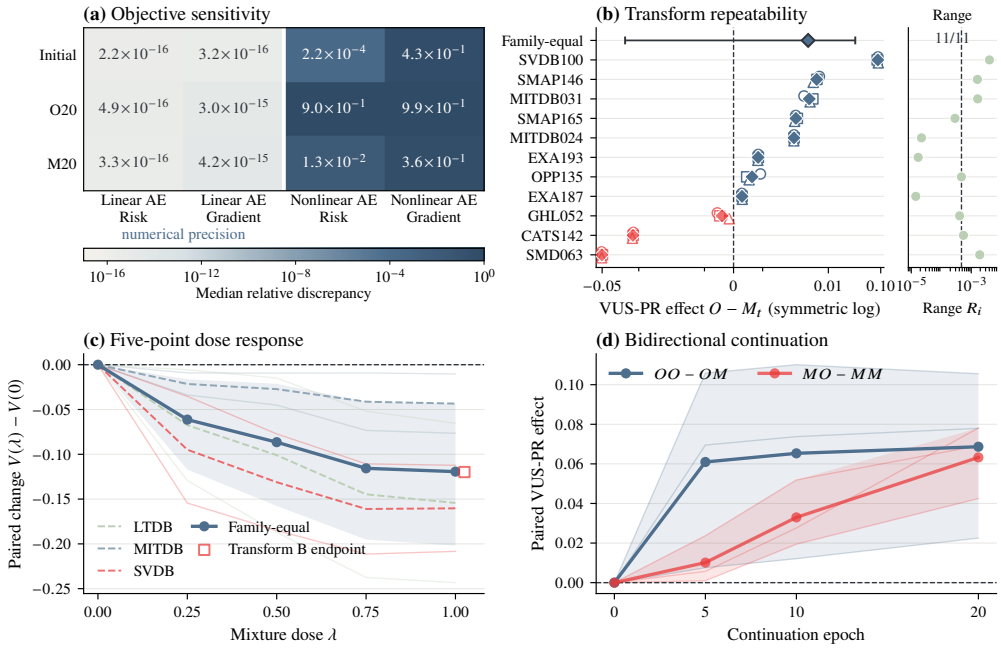


Figure 3: Effects of normal training information on learning and detection. O/M denote original/moment-matched windows; A/B/C are independent transforms. O20/M20 denote weights after 20 epochs, and paired letters indicate initial and continuation data. (a) Relative risk and gradient discrepancies at initialization and O20/M20. Each cell summarizes three sequences and three seeds by their median. (b) $O - M_t$ effects across 11 sequences. Circles/squares/triangles denote A/B/C, diamonds their means, and segments/sidebar their ranges. The symmetric-log axis has a linear center of ± 0.002 and affects display only. (c) Changes from $\lambda = 0$ across six sequences. Thin/dashed/thick lines show sequence/family/family-equal means; the open square is the transform-B endpoint. (d) $D_O = V_{OO} - V_{OM}$ and $D_M = V_{MO} - V_{MM}$ across three sequences, with V denoting VUS-PR. Pale/thick lines show sequence/family-equal means. The family-equal summary in (b) and bands in (c–d) use 95% paired bootstrap intervals; resampling details are in Appendix C.1.

4.3 Score Selectivity Within the Global Subspace

To isolate score variation beyond PCA residual, Figure 4(a) restricts queries to the same rank-16 PCA subspace or adds an identical off-subspace residual to both query groups. These two controls

Table 1: Key paired training-intervention comparisons. Differences first average three seeds within each sequence, then average within families and weight families equally. Positive sequences counts sequence-mean differences greater than zero. Complete intervals are in Appendices C.2–C.4.

Comparison	Sequences/Families	Δ VUS-PR	Positive sequences
Mechanism: Original − Matched	6/3	+0.119	6/6
Extension: $O - \frac{1}{3}(M_A + M_B + M_C)$	11/8	+0.004	8/11
Order control: Reordered − Original	6/3	+0.001	2/6
Deterministic full-batch: $O - M$	3/3	+0.076	3/3
Original-parent retention: OO − OM	3/3	+0.069	3/3
Matched-parent restoration: MO − MM	3/3	+0.063	3/3

Table 2: Real anomaly events with approximately matched PCA scores. Win rates count higher anomaly scores as wins and ties as 0.50, averaging pairs and seeds within events, then weighting matchable events equally within each family.

Family	Matched/Total events	ShallowAE win	PCA win	Difference
MITDB	3/3	0.919	0.658	+0.261
LTDB	26/122	1.000	0.346	+0.654
SVDB	54/80	0.827	0.463	+0.364
SMAP	1/2	0.778	0.333	+0.444
SWaT	2/6	0.056	0.500	−0.444
Exathlon	2/7	1.000	1.000	0.000

exactly preserve coefficient means and covariances; the third adds empirical marginal constraints with approximate moment matching. PCA scores are identical to numerical precision in all three controls, while the AUROC for separating original from surrogate combinations using ShallowAE scores is 0.903–0.996. ShallowAE scores therefore also depend on variable combinations within the retained subspace. Construction errors and replicate counts are in Appendix D.1.

Real-event matching tests the corresponding phenomenon in natural anomalies. Pairing anomalous and normal windows with approximately equal PCA scores covers 88/223 events. Among matchable events, ShallowAE has higher win rates than PCA in MITDB, LTDB, SVDB, and SMAP, a lower rate in SWaT, and the same rate in Exathlon. Appendix D.2 reports complete results and coverage.

4.4 Conditional relations shape reconstruction and scoring

Interventions on real data can change multiple joint statistics simultaneously. The theoretical P_+, P_- construction instead preserves coordinate-wise marginals, means, and covariance, isolating a specific relation through opposite conditional means. Figure 4(b) branches from the same P_+ parent model: retaining the relation keeps the residual valley along it, whereas switching to P_- redirects the shared-scale score field and reconstructions of fixed inputs toward the other relation.

In Figure 4(c), reconstruction-relation preference and target-task AUROC change together during continuation in both directions, with final full-test AUROC averaged over three seeds exceeding **0.98** in each direction. Score preferences on fixed queries reverse in both directions, showing that changes in reconstruction-relation preference accompany changes in anomaly rankings. The finite network thereby follows the direction predicted by conditional-mean theory: changing the normal relation redirects reconstruction geometry and reverses score preferences. Exact endpoints and clear-region results are in Appendix E.

4.5 Discussion

Learning normal relations can shape the reconstruction map so that anomaly scores reflect an input's compatibility with those relations. In the relation-switching experiment, reversing the training relation changes score preferences for the same inputs while preserving coordinate-wise marginals and covariance, showing that this compatibility depends on the normal training environment. Its

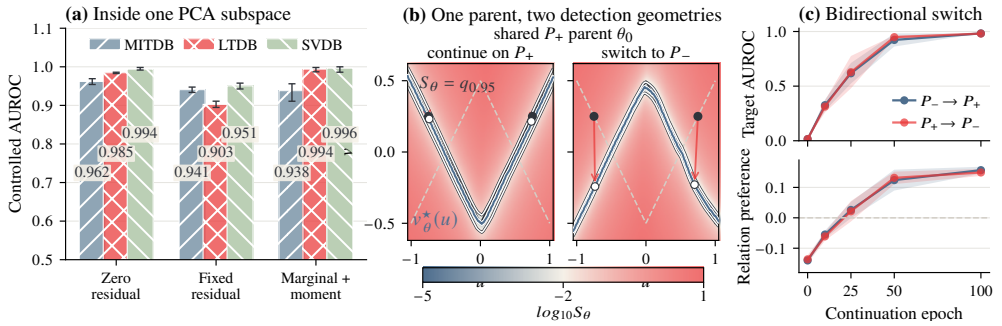


Figure 4: From within-subspace score selectivity to relation switching. (a) Bars show AUROC − 0.50, with raw AUROC inside and min–max whiskers over seed summaries or qualified replicates; counts are in Appendix D.1. (b) Two branches from the same P_+ parent (seed 2026) continue on P_+ or P_-, sharing a $\log_{10} S_\theta$ scale. Blue curves mark column-wise minima on the finite grid. Contours $S_\theta = q_{0.95}$ use each model's target-environment calibration and thus reflect both scores and thresholds. Solid/dashed lines denote target/alternative relations; black/white points denote identical queries/their reconstructions. (c) Relation preference is the output's squared distance to the alternative relation minus that to the target; positive values favor the target. Both trajectories show the mean and one standard deviation over three seeds. Calibration and grid details are in Appendix E.

detection value further depends on where normal and anomalous inputs lie relative to the learned relation: under a uniform reconstruction approximation bound on the query domain, anomalous deviations exceeding the normal-relation width plus twice the approximation error suffice for strict score separation. The collinear-gap experiment (Appendix F) further shows that unseen inputs along the extension of a learned relation can also receive low scores. Understanding detection behavior therefore requires examining how the reconstruction map extends beyond the training samples.

Negative effects on real data show that selectivity over variable combinations can also reduce separation between normal and anomalous inputs. In Figure 3(b), moment-matched training performs better on GHL052, CATS142, and SMD063. One possible explanation is that some relations in the original training data are specific to the training period, and altering these constraints reduces high-score responses to new normal combinations. The two matchable SWaT events in Table 2 illustrate another case: among window pairs with approximately equal PCA scores, ShallowAE more often assigns lower scores to anomalous windows. These results indicate that additional selectivity over combinations can raise anomaly scores, but can also raise scores for normal inputs or preserve low scores for anomalies.

The difference between PCA and nonlinear reconstruction can thus be understood in terms of their selectivity over combinations within the retained subspace. PCA assigns zero residual to every input in that subspace, accommodating normal variation along it while leaving anomalies that violate relations within it difficult to detect. ShallowAE can assign different scores to these combinations; whether this helps detection depends on whether the learned relations remain applicable in the test environment and whether anomalies deviate from them. A promising direction for future research is selective adaptation of normal relations: identifying relations that remain valid across operating states and using them as stable criteria for anomaly detection.

5 Conclusion

We characterize the statistical sufficiency boundary of squared-reconstruction objectives. Controlled experiments show that normal training information beyond global second-order statistics can change ShallowAE's learning and detection behavior, while relation switching shows how normal-variable combinations shape low-residual regions and fixed-input scores. These results support relation-sensitive reconstruction as one source of score selectivity beyond global linear-subspace distance; detection gains require learned relations that remain valid for normal test variation and anomalies that depart from them.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and the identification of potentially relevant literature. The authors independently verified the relevant search results and AI-assisted revisions and take full responsibility for this paper.

ETHICS STATEMENT

This study uses existing benchmark time series and synthetic data to examine detection behavior. Clinical and operational decision validity lies outside the scope of these experiments.

REPRODUCIBILITY STATEMENT

The appendices provide theoretical proofs, evaluation conventions, training configurations, and numerical consistency checks. They specify the statistical units and aggregation procedures used for the reported results. Code and supplementary implementation materials are available in the anonymous repository at `https://anonymous.4open.science/r/BeyondCov-858F`.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(110):3743–3773, 2014.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3395–3404, 2020.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, 132:108945, 2022.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Randall Balestriero and Richard G. Baraniuk. A spline theory of deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 374–383, 2018.

Paul Boniol, Ashwin K. Krishna, Marine Bruel, Qinghua Liu, Mingyi Huang, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, Michael J. Franklin, and John Paparrizos. VUS: Effective and efficient accuracy measures for time-series anomaly detection. *The VLDB Journal*, 34(3):32, 2025.

José Camacho, Alejandro Pérez-Villegas, Pedro García-Teodoro, and Gabriel Maciá-Fernández. PCA-based multivariate statistical network monitoring for anomaly detection. *Computers & Security*, 59:118–137, 2016.

Yanbo Chen and Weiwei Liu. Two-layer convolutional autoencoders trained on normal data provably detect unseen anomalies. In *International Conference on Learning Representations*, 2026.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

Utsav Dutta, Gerardo Pastrana, Sina Khoshfetrat Pakazad, and Henrik Ohlsson. Giving sensors a voice: Multimodal JEPA for semantic time-series embeddings. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

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

Kamil Faber, Marcin Pietroń, Dominik Żurek, and Roberto Corizzo. xLSTMAD: A powerful xlstm-based method for anomaly detection. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 247–256, 2025.

Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

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

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

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108231–108261, 2024.

Kadir-Kaan Özer, René Ebeling, and Markus Enzweiler. STREAM-VAE: Dual-path routing for slow and fast dynamics in vehicle telemetry anomaly detection. In *2026 IEEE Intelligent Vehicles Symposium (IV)*, pp. 444–451, 2026.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):38:1–38:38, 2021.

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

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3–4):346–382, 2000.

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

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1–4):77–94, 1992.

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

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*, 2023.

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

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 2018 World Wide Web Conference*, pp. 187–196, 2018.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *International Conference on Learning Representations*, pp. 26295–26318, 2024.

Chin-Chia Michael Yeh. Matrix profile for time-series anomaly detection: A reproducible open-source benchmark on TSB-AD, 2026. URL `https://arxiv.org/abs/2604.02445`.

Sangwoong Yoon, Yung-Kyun Noh, and Frank C. Park. Autoencoding under normalization constraints. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12087–12097, 2021.

Zhijie Zhong, Zhiwen Yu, Xing Xi, Yue Xu, Wenming Cao, Yiyuan Yang, Kaixiang Yang, and Jane You. SimAD: A simple dissimilarity-based approach for time-series anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19669–19680, 2025.

Quan Zhou, Changhua Pei, Fei Sun, Han Jing, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. KAN-AD: Time series anomaly detection with kolmogorov–arnold networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 79136–79149, 2025.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43322–43355, 2023.

A Theoretical Proofs and Analytical Boundaries

A.1 Characterization of second-moment sufficiency

Let $\Pi_2(\Omega)$ denote the space of polynomials of degree at most two on $\Omega \subseteq \mathbb{R}^D$.

For the forward direction, if $S_\theta \in \Pi_2(\Omega)$, there exist a scalar c_0, vector β, and matrix B such that

$$S_\theta(x) = c_0 + \beta^\top x + \mathrm{tr}(Bxx^\top). \tag{A.1}$$

The antisymmetric part of B does not contribute to the quadratic form, so B may be taken symmetric. Hence

$$\mathbb{E}_P S_\theta(X) = c_0 + \beta^\top \mathbb{E}_P[X] + \mathrm{tr}\big(B\mathbb{E}_P[XX^\top]\big)\,, \tag{A.2}$$

which is identical under P and P' whenever $P \equiv_2 P'$.

For the converse, because $S_\theta \notin \Pi_2(\Omega)$, one can select finitely many points $x_1, \ldots, x_m \in \Omega$ such that

$$s_{\mathrm{val}} = (S_\theta(x_1), \ldots, S_\theta(x_m))^\top \tag{A.3}$$

does not belong to the value subspace induced by $\Pi_2(\Omega)$ on those points. Otherwise, fitting a quadratic polynomial on a unisolvent set and adding arbitrary points of Ω one by one would force S_θ to equal the same quadratic polynomial throughout Ω, contradicting the assumption.

Finite-dimensional linear algebra therefore gives a nonzero vector $c = (c_1, \ldots, c_m)^\top$ orthogonal to every quadratic-polynomial value vector but not to s_{val}:

$$\sum_{i=1}^m c_i p(x_i) = 0 \quad \text{for every } p \in \Pi_2(\Omega), \qquad \sum_{i=1}^m c_i S_\theta(x_i) \neq 0. \tag{A.4}$$

Because constants belong to Π_2, $\sum_i c_i = 0$. Define

$$c_i^+ = \max(c_i, 0), \qquad c_i^- = \max(-c_i, 0), \qquad M = \sum_i c_i^+ = \sum_i c_i^- > 0, \tag{A.5}$$

and finite-support probability distributions

$$P = \sum_i \frac{c_i^+}{M} \delta_{x_i}, \qquad P' = \sum_i \frac{c_i^-}{M} \delta_{x_i}. \tag{A.6}$$

Here δ_{x_i} is the Dirac probability measure at x_i. Taking $p(x)$ to be each coordinate x_j and each product $x_j x_k$ yields

$$\mathbb{E}_P[X] = \mathbb{E}_{P'}[X], \qquad \mathbb{E}_P[XX^\top] = \mathbb{E}_{P'}[XX^\top]. \tag{A.7}$$

But nonorthogonality to the score vector gives $\mathbb{E}_P S_\theta(X) \neq \mathbb{E}_{P'} S_\theta(X)$. Both distributions have finite support and hence all required moments are finite, proving Theorem 1.

A.2 Affine risk and PCA corollaries

Corollary 1.1 (Affine reconstruction-objective invariance)**.** *If $r_\theta(x) = A_\theta x + b_\theta$ and P has mean μ and covariance Σ, its per-coordinate population risk is*

$$L_P(\theta) = \frac{1}{D}\left[\operatorname{tr}\big((I - A_\theta)\Sigma(I - A_\theta)^\top\big) + \|(I - A_\theta)\mu - b_\theta\|_2^2\right]. \tag{A.8}$$

Moment-matched distributions have the same affine risk at every parameter value and the same population gradients wherever differentiable; empirical moment matching likewise preserves empirical full-batch gradients.

For $r_\theta(x) = A_\theta x + b_\theta$, write $X = \mu + X_c$ with

$$\mathbb{E}[X_c] = 0, \qquad \mathbb{E}[X_c X_c^\top] = \Sigma. \tag{A.9}$$

The residual is

$$X - r_\theta(X) = (I - A_\theta)X_c + (I - A_\theta)\mu - b_\theta. \tag{A.10}$$

The random–constant cross term has zero expectation, and

$$\mathbb{E}\|(I - A_\theta)X_c\|_2^2 = \operatorname{tr}\big((I - A_\theta)\Sigma(I - A_\theta)^\top\big). \tag{A.11}$$

Dividing by D proves Corollary 1.1, including gradient invariance wherever differentiable. A common differentiable parameter regularizer preserves this equality.

For the empirical transform, let $H_n = I_n - 1_n 1_n^\top / n$. Orthogonality and $Q1_n = 1_n$ imply $Q^\top 1_n = 1_n$ and $H_n Q = Q H_n$. Thus

$$1_n^\top Q X_{\mathrm{N}} = 1_n^\top X_{\mathrm{N}}, \qquad (H_n Q X_{\mathrm{N}})^\top (H_n Q X_{\mathrm{N}}) = (H_n X_{\mathrm{N}})^\top (H_n X_{\mathrm{N}}), \tag{A.12}$$

so the empirical mean and centered Gram matrix are unchanged under either common covariance denominator.

Corollary 1.2 (Fixed-PCA-subspace control)**.** *Let $U_r^\top U_r = I_r$, $U_r^\top v_\perp = 0$, and $y = \mu + U_r\alpha + v_\perp$. Projection onto $\mu + \operatorname{span}(U_r)$ has squared residual $S_{\mathrm{PCA}}(y) = \|v_\perp\|_2^2$, independently of the coefficient combination α. With $v_\perp = 0$, every query has zero PCA residual. Variation in ShallowAE scores across these queries therefore reflects selectivity within the global linear support.*

Under the assumptions of Corollary 1.2, set $P_r = U_r U_r^\top$. Then

$$(I - P_r)(y - \mu) = (I - U_r U_r^\top)(U_r\alpha + v_\perp) = v_\perp, \tag{A.13}$$

and hence $S_{\mathrm{PCA}}(y) = \|v_\perp\|_2^2$. With fixed $v_\perp$, coefficient distributions with the same mean and covariance induce inputs with the same mean and covariance:

$$\mathbb{E}[y] = \mu + U_r\mathbb{E}[\alpha] + v_\perp, \qquad \operatorname{Cov}(y) = U_r \operatorname{Cov}(\alpha) U_r^\top. \tag{A.14}$$

This proves Corollary 1.2 and the finite-sample invariance used by the controlled experiment.

A.3 SUFFICIENCY OF ACTIVATION-CONDITIONED MOMENTS

On a ReLU activation region $\mathcal{R}_s$, write

$$r_\theta(x) = A_s x + b_s, \qquad C_s = I - A_s. \tag{A.15}$$

The regional contribution to expected squared residual is

$$\mathbb{E}\left[\|C_s X - b_s\|_2^2 1_{\mathcal{R}_s}(X)\right]. \tag{A.16}$$

Expanding gives

$$\mathbb{E}[X^\top C_s^\top C_s X 1_{\mathcal{R}_s}(X)] - 2b_s^\top C_s \mathbb{E}[X 1_{\mathcal{R}_s}(X)] + \|b_s\|_2^2 P(X \in \mathcal{R}_s). \tag{A.17}$$

Using

$$\mathbb{E}[X^\top C_s^\top C_s X 1_{\mathcal{R}_s}(X)] = \operatorname{tr}\Big(C_s M_s^{(2)} C_s^\top\Big) \tag{A.18}$$

and summing over regions proves Equation 4. Zero-mass regions contribute zero. Because a continuous ReLU network agrees on shared boundaries, any consistent measurable assignment of boundaries gives the same risk.

For the local gradient claim, consider a finite training set whose hidden-unit preactivations are all nonzero at parameters θ_0. Continuity and finiteness give a common parameter neighborhood in which none of these signs changes. In that neighborhood each sample uses a fixed affine expression in the input, so the empirical risk is the corresponding finite sum of regional quadratic forms and its parameter derivatives are determined by the empirical regional moments together with the model parameters. The statement is local: after a boundary crossing, samples may change regions and the decomposition must be updated.

A.4 CONDITIONAL-EXPECTATION PROJECTION LEMMA

Fix $Z = f^{(e)}(X)$ with $X \in L^2$, and let $g_P^*(Z) = \mathbb{E}[X \mid Z]$. For any square-integrable measurable decoder g, expand $X - g(Z) = (X - g_P^*(Z)) + (g_P^*(Z) - g(Z))$. The cross term vanishes after conditioning on Z, since $\mathbb{E}[X - g_P^*(Z) \mid Z] = 0$. Therefore

$$\mathbb{E}\|X - g(Z)\|_2^2 = \mathbb{E}\|X - g_P^*(Z)\|_2^2 + \mathbb{E}\|g(Z) - g_P^*(Z)\|_2^2. \tag{A.19}$$

The final term is nonnegative and vanishes exactly when $g(Z) = g_P^*(Z)$ almost surely, proving Lemma 3. The lemma characterizes optimality for a fixed representation; convergence of a jointly trained finite encoder–decoder requires separate analysis.

A.5 IDENTICAL STATISTICS AND OPPOSITE RELATIONS

Let $W = h(U) = |U| - 1/2$. Because $|U| \sim \mathrm{Unif}[0,1]$,

$$W \sim \mathrm{Unif}[-1/2, 1/2], \qquad \mathbb{E}[W] = 0, \qquad \mathbb{E}[W^2] = \frac{1}{12}. \tag{A.20}$$

Since W and $-W$ have the same distribution and η is independent and symmetric, $V_+ = W + \eta$ and $V_- = -W + \eta$ have identical marginals. Both U and V_κ have mean zero.

Moreover, the first integrand below is odd, while the second term vanishes by independence:

$$\mathbb{E}[U V_\kappa] = \kappa\, \mathbb{E}[U(|U| - 1/2)] + \mathbb{E}[U\eta] = 0. \tag{A.21}$$

Thus, $\mathrm{Cov}(U, V_\kappa) = 0$. Furthermore,

$$\mathrm{Var}(V_\kappa) = \mathrm{Var}(W) + \mathrm{Var}(\eta) = \frac{1}{12} + \frac{\delta^2}{3}. \tag{A.22}$$

Together with $\mathrm{Var}(U) = 1/3$, this gives the shared covariance in Theorem 4:

$$\mathrm{Cov}\big((U, V_\kappa)^\top\big) = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/12 + \delta^2/3 \end{pmatrix}. \tag{A.23}$$

Conditioning on $U = u$ gives

$$\mathbb{E}[V_\kappa \mid U = u] = \kappa\, h(u) + \mathbb{E}[\eta] = \kappa\, h(u). \tag{A.24}$$

For $r_m(u, v) = (u, m(u))$, the population risk is

$$R_{P_\kappa}(m) = \mathbb{E}[(V_\kappa - m(U))^2]. \tag{A.25}$$

Applying Lemma 3 with representation U yields the L^2-optimal relation $m^*_\kappa(u) = \kappa h(u)$, unique almost everywhere under the law of U, and

$$R_{P_\kappa}(m^*_\kappa) = \mathbb{E}[\eta^2] = \frac{\delta^2}{3}. \tag{A.26}$$

Using the opposite relation under $P_{-\kappa}$ gives residual

$$V_{-\kappa} - m^*_\kappa(U) = -2\kappa\, h(U) + \eta. \tag{A.27}$$

The cross term vanishes by independence and centering, so

$$R_{P_{-\kappa}}(m^*_\kappa) = 4\mathbb{E}[h(U)^2] + \mathbb{E}[\eta^2] = \frac{1}{3} + \frac{\delta^2}{3}. \tag{A.28}$$

Subtracting the optimal noise risk gives the exact excess risk $1/3$.

Finally, because $1/3 > 1/12 + \delta^2/3$ for the experimental value $\delta = 0.03$, rank-1 PCA uniquely retains the U axis in both environments. Its residual is V_κ. Since V_+ and V_- have the same distribution, the PCA score V^2_κ also has the same distribution. The joint distributions remain different, while PCA score invariance follows from the matched first- and second-order statistics used by PCA.

The finite training grid differs slightly from the population distribution: midpoint discretization gives $\mathbb{E}[W^2] = (1-128^{-2})/12$, and noise discretization gives $\mathbb{E}[\eta^2] = \delta^2(1-8^{-2})/3$. The finite-sample means, marginals, and covariance nevertheless match exactly between environments.

A.6 Relation approximation and score separation

Write $\Delta r_\theta = r_\theta - r_*$. Applying the triangle and reverse-triangle inequalities to $x - r_\theta(x) = (x - r_*(x)) - \Delta r_\theta(x)$ gives

$$\|x - r_*(x)\|_2 - \|\Delta r_\theta(x)\|_2 \le \|x - r_\theta(x)\|_2 \le \|x - r_*(x)\|_2 + \|\Delta r_\theta(x)\|_2. \tag{A.29}$$

Using $\|\Delta r_\theta(x)\| \le \varepsilon$, nonnegativity, and $\|x - r_*(x)\|_2 = d_f(x)$ proves Equation 10.

If $d_f(x_0) \le \delta$ for every normal input and $d_f(x_1) \ge a$ for every relation-violating input, then

$$\sqrt{S_\theta(x_0)} \le \delta + \varepsilon, \qquad \sqrt{S_\theta(x_1)} \ge a - \varepsilon. \tag{A.30}$$

The condition $a > \delta + 2\varepsilon$ yields a strict gap. For the opposite-relation construction, an alternative-environment input at a fixed u differs from the target center relation by at least $2|h(u)| - \delta$. On $|h(u)| \ge b$,

$$\sqrt{S_{\text{alternative}}} \ge 2b - \delta - \varepsilon, \qquad \sqrt{S_{\text{normal}}} \le \delta + \varepsilon, \tag{A.31}$$

so $b > \delta + \varepsilon$ suffices for strict separation. This proves Corollary 5. The uniform approximation assumption concerns the full query domain; a finite-grid maximum characterizes approximation only on the sampled grid.

A.7 Analytical boundaries and path identity

For a path $x(t)$ joining two normal points, define the residual barrier

$$B_{\text{path}} = \max_{t\in[0,1]} S_\theta(x(t)) - \max\{S_\theta(x(0)), S_\theta(x(1))\}. \tag{A.32}$$

A large barrier means that the endpoints are easy to reconstruct while some intervening combinations have higher residuals. This measures score-landscape shape; anomaly separation also depends on the normal-score background and is therefore evaluated relative to a normal-calibration threshold.

For the bent relation, let $0 < \ell \le 1$ and

$$r(u, v) = (u, |u| - 1/2), \qquad x(t) = \big(\ell(2t-1), \ell - 1/2\big). \tag{A.33}$$

The first coordinate reconstructs exactly, while the second residual is

$$\ell - \frac{1}{2} - \left(|\ell(2t-1)| - \frac{1}{2}\right) = \ell\big(1 - |2t-1|\big). \tag{A.34}$$

Hence

$$S(x(t)) = \ell^2\big(1 - |2t-1|\big)^2. \tag{A.35}$$

Both endpoints have score zero and the midpoint has score ℓ^2, giving a residual barrier between two reconstructable points. If the normal band contains only vertical noise of absolute value at most δ, its normal score is at most δ^2.

Conversely, $r(u, v) = (u, 0)$ reconstructs every point on the line $v = 0$, including any interval omitted from the training support. This collinear-gap example shows that the low-residual set can strictly contain the normal support and that absence from the observed support does not itself force high reconstruction error.

For the path identity, define the local residual-response operator $K_\theta(x) = I - J_{r_\theta}(x)$, where J_{r_θ} is the reconstruction Jacobian. A finite ReLU network is globally Lipschitz. Its composition with an absolutely continuous path is absolutely continuous, as is the squared norm of the residual on the bounded path image. At almost every point where r_θ is differentiable, the chain rule gives

$$\begin{aligned} \frac{d}{dt} S_\theta(x(t)) &= 2\big(x(t) - r_\theta(x(t))\big)^\top \big(I - J_{r_\theta}(x(t))\big)\dot{x}(t) \\ &= 2e_\theta(x(t))^\top K_\theta(x(t))\dot{x}(t). \end{aligned} \tag{A.36}$$

Integration over $[0, 1]$ recovers the endpoint score difference. Activation-boundary points have measure zero for an ordinary piecewise-smooth transversal path; if a path lies on a boundary for an interval, continuity makes adjacent affine maps agree on that face and the tangential derivative of the restricted map gives the same identity. This proves Equation A.36 without requiring an arbitrary full Jacobian selection on the boundary.

B Complete Benchmark and Parameter Sensitivity

B.1 Complete TSB-AD-M ranking

Table B.1 summarizes 35 methods on 180 multivariate time series from 17 dataset families. The benchmark datasets and associated resources are available from the official TSB-AD project page.[1] ShallowAE and GWPCA are our implementations. For CATCH (Wu et al., 2025), DADA (Shentu et al., 2025), KAN-AD (Zhou et al., 2025), PaAno (Park & Kang, 2026), and SimAD (Zhong et al., 2025), we adapted the authors' official implementations to the TSB-AD data-loading, training, and scoring interfaces.[2] Sequence VUS-PR is the arithmetic mean over equally weighted sequences; Rank orders methods by this unrounded mean, with exact ties receiving competition ranks. Average rank averages the within-sequence ranks, with tied scores assigned their average position. Above mean counts sequences on which a method strictly exceeds the mean score of all 35 methods. Bottom-5 counts sequences on which it scores at or below the fifth-lowest method score, including all cutoff ties. Both counts use a denominator of 180; Bottom-5 measures relative standing within each sequence.

[1] TSB-AD project page.

[2] Official repositories: CATCH, DADA, KANAD, PaAno, and SimAD.

Table B.1: Complete TSB-AD-M ranking by sequence-averaged VUS-PR, with per-sequence rank and count summaries. Higher VUS-PR and Above mean are better; lower Average rank and Bottom-5 are better.

Rank	Method	VUS-PR	Average rank	Above mean	Bottom-5/180
1	ShallowAE	**0.456**	11.192	110/180	9/180
2	GWPCA	0.448	11.750	106/180	9/180
3	STREAM-VAE (2026)	0.435	11.108	106/180	5/180
4	PaAno (2026)	0.420	14.228	104/180	15/180
5	xLSTMAD (2025)	0.392	12.167	121/180	12/180
6	CHARM (2026)	0.391	13.600	105/180	23/180
7	TSPulse (2026)	0.387	15.978	106/180	34/180
8	KAN-AD (2025)	0.367	12.144	96/180	5/180
9	CATCH (2025)	0.356	15.494	91/180	26/180
10	MMPAD (2026)	0.354	16.717	82/180	34/180
11	SimAD (2025)	0.316	15.428	67/180	4/180
12	CNN	0.313	13.983	85/180	4/180
13	OmniAnomaly (2019)	0.312	16.958	82/180	36/180
14	LSTMAD	0.307	15.239	77/180	13/180
15	USAD (2020)	0.304	17.397	81/180	38/180
16	DADA (2025)	0.302	18.311	61/180	16/180
17	AutoEncoder	0.295	17.542	63/180	23/180
18	KMeansAD	0.295	14.517	86/180	18/180
19	PCA	0.275	18.356	61/180	33/180
20	CBLOF	0.273	18.364	59/180	21/180
21	MCD	0.271	18.806	49/180	26/180
22	OCSVM	0.265	21.053	56/180	35/180
23	Donut (2018)	0.260	18.111	55/180	21/180
24	RobustPCA	0.239	21.503	49/180	50/180
25	FITS (2024)	0.215	23.711	21/180	36/180
26	OFA (2023)	0.211	23.156	16/180	39/180
27	EIF	0.210	19.886	42/180	23/180
28	COPOD	0.203	21.669	38/180	30/180
29	IForest	0.203	21.592	39/180	22/180
30	HBOS	0.190	22.064	28/180	28/180
31	TimesNet (2023)	0.185	24.261	12/180	36/180
32	KNN	0.184	24.136	23/180	43/180
33	TranAD (2022)	0.176	20.189	39/180	37/180
34	LOF	0.139	24.683	15/180	47/180
35	AnoTrans (2022)	0.118	24.708	15/180	62/180

B.2 Parameter sensitivity

Figure B.1 reports a unified scan on all 180 TSB-AD-M sequences, with every model trained from scratch using seed 2026. We vary activation, window length, and hidden width individually around the reference configuration $(L, H, d) = (100, 64, 16)$ with ReLU and 20 epochs; bottlenecks $d \in \{4, 8, 16, 32\}$ are each trained for 100 epochs and evaluated at epochs 1, 2, 5, 10, 20, 40, 60, and 100. All configurations use Adam with learning rate 0.001, weight decay 0.00001, and batch size 128. Scores are evaluated on the full series using a fixed VUS tolerance window of 100, independently of model window length. We report sequence-equal means. The benchmark-designated training prefixes are retained without label-based filtering; 20 of the 180 prefixes contain labeled anomalies. This scan therefore evaluates configuration sensitivity under the benchmark's native, partly contaminated training convention. Its evaluation protocol is reported separately from the benchmark comparison.

At 20 epochs, ReLU achieves a mean VUS-PR of 0.470, compared with 0.450 for linear activation; Leaky-ReLU and GELU reach 0.466 and 0.465. Window length has a larger effect on the mean than hidden width in the tested ranges: increasing L from 25 to 100 raises VUS-PR from 0.387 to

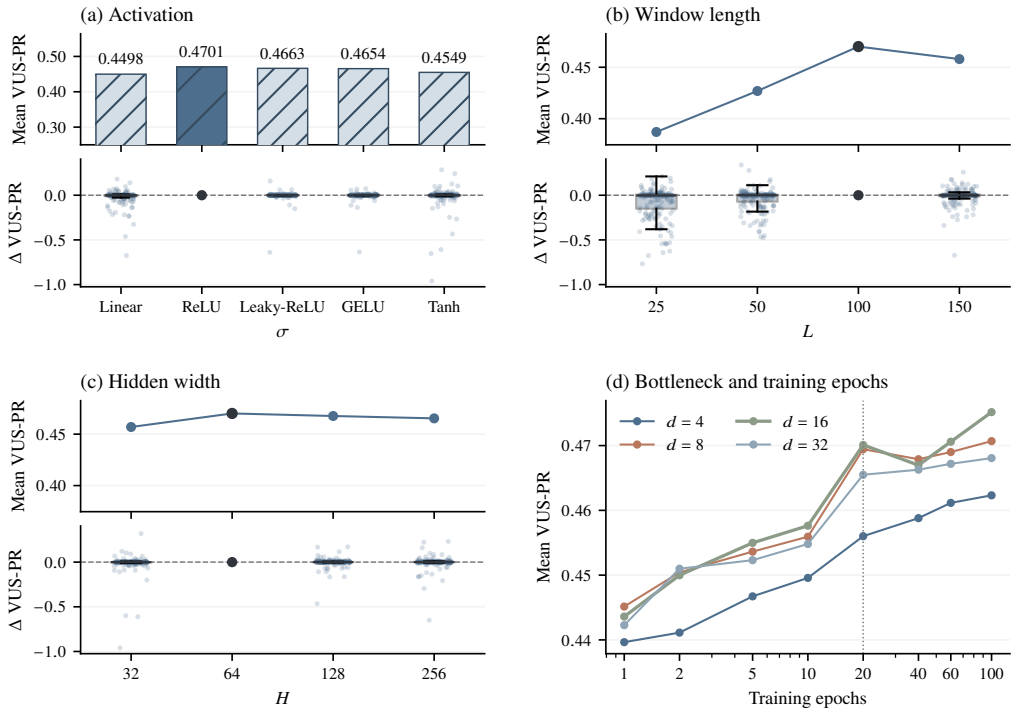


Figure B.1: ShallowAE parameter sensitivity on 180 sequences with seed 2026. Upper plots in (a–c) show sequence-equal mean VUS-PR for activation, window length, and hidden width; the zero-baseline bars in (a) are labeled with their means. Dark marks identify the reference configuration. Lower plots show per-sequence changes relative to that configuration: boxes denote interquartile ranges, center lines medians, whiskers extend to observations within 1.5 interquartile ranges, and dots show all sequences. These distributions summarize variation across sequences; the reference has zero change. (d) Mean VUS-PR across training epochs for four bottleneck dimensions, with a logarithmic epoch axis and a reference line at 20 epochs.

0.470, whereas $L = 150$ gives 0.458. Increasing hidden width beyond 64 does not improve the mean. The paired distributions show how these aggregate changes vary across individual sequences. The highest observed mean, 0.475, occurs at $L = 100$, $H = 64$, $d = 16$, ReLU, and 100 epochs, an increase of 0.005 over the 20-epoch reference. This value is the descriptive maximum within the tested configurations. The scan characterizes configuration sensitivity at seed 2026; initialization robustness and causal tests of the conditional-relation mechanism require separate experiments.

C Training Intervention Evidence

C.1 Common protocol and objective sensitivity

We distinguish three evidence sets. The original mechanism set contains MITDB019/020, LTDB079/083, and SVDB084/085 (six sequences, three families). Same-weight, full-batch, and continuation controls use MITDB019, LTDB079, and SVDB084. The cross-family extension contains 11 different sequences from eight families. These sets are analyzed separately.

Except for the two-dimensional toys, windows have length 100 and are standardized channel-wise and flattened in channel-major order. ShallowAE uses hidden width 64, bottleneck 16, ReLU, and seeds 2026/2027/2028. Normal-prefix blocks for training, reference, calibration, and probing have proportions 50%/20%/15%/15%, separated by 99 timestamps to prevent window overlap. Training uses Adam, learning rate 10^{-3}, batch size 128, weight decay 10^{-5}, and 20 epochs, with deterministic CUDA algorithms. Full-batch and continuation budgets are specified below.

Statistical units. Paired effects first average neural seeds within sequences, then sequences within families, and finally weight families equally. Ordinary-training, cross-transform, and dose-response intervals use 10,000 paired bootstrap draws over families and then sequences within each sampled family. Full-batch and continuation controls have one sequence per family, so the procedure reduces to family-level resampling. Families and sequences provide the dataset-level replication; windows, path nodes, and neural seeds serve as within-dataset samples or repeated fits. Figure 3(b) shows transform min–max ranges for individual sequences; its aggregate error bar and the bands in (c–d) are bootstrap intervals. Figure 4(a) uses min–max ranges as specified in Appendix D.1; panel (c) uses three-seed means and standard deviations. Event win rates average pairs and seeds within events, then weight matchable events equally within families.

For $X_{\mathrm{N}} \in \mathbb{R}^{n\times D}$, the matched construction is $\widetilde{X}_{\mathrm{N}} = QX_{\mathrm{N}}$ with $Q^\top Q = I_n$ and $Q1_n = 1_n$. Across the six sequences, maximum mean and covariance errors are 5.46×10^{-17} and 8.52×10^{-16}, respectively; the maximum covariance error after float32 conversion is 5.82×10^{-9}. The maximum exact-PCA projector discrepancy is 5.12×10^{-14}, and the VUS-PR change is 0.00000. These numerical-integrity checks are distinct from the model risk and gradient discrepancies below.

Objective and full gradient at identical weights. Let L_O, L_M be per-coordinate reconstruction MSE on the complete original and matched training sets, and let g_O, g_M concatenate the corresponding gradients over all model parameters at the same weights. The audited relative discrepancies are

$$d_L = \frac{|L_O - L_M|}{\max(|L_O|, |L_M|, 10^{-30})}, \qquad d_g = \frac{\|g_O - g_M\|_2}{\max(\|g_O\|_2, \|g_M\|_2, 10^{-30})}. \tag{C.1}$$

Figure 3(a) uses float64 exactly matched inputs; each cell is the median over three sequences and three seeds at initialization, O20, or M20. O20/M20 denote weights after 20 epochs on original/matched data. Across all weight states, maximum linear-AE risk and gradient discrepancies are 8.59×10^{-16} and 6.74×10^{-15}; nonlinear-AE gradient discrepancies range from 0.237 to 1.004, with a pooled median of 0.467. This pooled median combines all 27 records across states, while sequences and families remain the independent dataset units. These comparisons measure objective differences at the same parameter point, with training-run comparisons reported separately.

C.2 Training order and deterministic full-batch controls

Original, reordered, and matched training. Table C.1 reports ordinary training on six sequences from three families. Panel A gives sequence-level scores; Panel B reports family-equal paired contrasts after averaging the three neural seeds within each sequence.

Table C.1: Ordinary training controls. Panel A reports VUS-PR averaged over three neural seeds; Panel B reports family-equal paired differences and 95% confidence intervals.

Panel A — Sequence-level VUS-PR

Sequence	Original	Reordered	Matched
MITDB019	0.421	0.425	0.344
MITDB020	0.016	0.016	0.006
LTDB079	0.992	0.992	0.927
LTDB083	0.381	0.365	0.138
SVDB084	0.397	0.390	0.285
SVDB085	0.665	0.691	0.456

Panel B — Paired contrasts

Contrast	Δ VUS-PR	95% CI	Positive sequences
Original − Matched	+0.119	[0.043, 0.202]	6/6
Reordered − Original	+0.001	[−0.010, 0.015]	2/6

Note. Intervals resample families and then sequences within sampled families, using 10,000 paired bootstrap draws. The last column counts sequences whose mean difference is positive.

Deterministic full-batch optimization. From matched initializations, Original and Matched each receive 300 complete-training-set Adam updates. Table C.2 reports mean ± standard deviation over three neural seeds; differences first average seeds within sequences and then take Original minus Matched.

The interval comes from 10,000 family-level paired bootstrap draws and reflects the three evaluated families, each represented by one sequence.

Table C.2: Deterministic full-batch optimization with 300 Adam updates. Sequence scores are mean ± standard deviation over three neural seeds; the aggregate 95% CI uses 10,000 family-level paired bootstrap draws.

Sequence	Original	Matched	Δ VUS-PR
MITDB019	0.41179 ± 0.01061	0.35359 ± 0.02153	+0.05820
LTDB079	0.99215 ± 0.00069	0.92020 ± 0.00139	+0.07195
SVDB084	0.38195 ± 0.01091	0.28326 ± 0.00296	+0.09869
Family-equal mean	0.59530	0.51902	**+0.07628**

Note. $\Delta = \text{Original} - \text{Matched}$. The family-equal difference has 95% CI [0.05820, 0.09869], distinct from the seed standard deviations above.

The order control and deterministic full-batch experiment jointly rule out sample order and mini-batch noise as sufficient explanations.

C.3 Three-transform repeatability

The cross-family set was selected only from normal-prefix properties, without reading anomaly-detection scores. Selection excludes the 18 sequences previously used in covariance interventions. Eligibility requires at least 2,000 normal-prefix timestamps, no anomalous prefix labels, finite prefix inputs, the four-block split above with at least 32 windows per block, and at most 25% constant channels (prefix standard deviation $\leq 10^{-10}$). After channel-wise prefix standardization, define

$$d_{\text{shift}} = \sqrt{C^{-1}\|\bar{z}_{\text{first}} - \bar{z}_{\text{second}}\|_2^2}, \tag{C.2}$$

where the two means use the first and second halves of the prefix. Among the 89 unused basically eligible sequences, the 75th-percentile cutoff is 0.592. We retain sequences at or below this cutoff, select the lowest-drift sequence in each eligible family, then add the lowest-drift second candidates until reaching 12, with at most two per family. The compute-complete set contains 11 sequences from eight families: unfinished SVDB102 is excluded, and every completed sequence is retained regardless of effect direction. Because the benchmark had been inspected previously, we treat this set as an extension rather than an independent confirmation. Each sequence is evaluated under three independent matched transforms.

Transform C uses seed 20260921 and the same DCT non-DC permutation and sign-flip rule as A and B. The three transforms differ but preserve the training-window mean and complete Gram matrix.

All 11 sequences have the same sign under A/B/C, and the median effect range is 0.000472. Family-equal effects are +0.00413 [-0.01841, 0.03234] for $O - M_A$; +0.00401 [-0.01795, 0.03155] for $O - M_B$; +0.00420 [-0.01899, 0.03289] for $O - M_C$; and +0.00411 [-0.01822, 0.03264] for their mean. Each has 8/11 positive sequences and 5/8 positive families. Excluding GHL gives +0.00476 [-0.02051, 0.03691], with 8/10 positive sequences and 5/7 positive families; the interval still includes zero. Transform C's exact PCA projector and pointwise-score discrepancies are at most 1.99×10^{-14} and 8.73×10^{-11}.

Thus the three tested random realizations are highly consistent within a scene, while different scenes can have opposite natural-anomaly ranking effects. This supports cross-transform repeatability; drift, overfitting, and anomaly type remain candidate explanations for scene differences. In Figure 3(b), segments and the side column report transform ranges, while the family-equal summary row reports a hierarchical-bootstrap interval.

Table C.3: Three-transform repeatability.

Family	ID	$O - M_A$	$O - M_B$	$O - M_C$	Mean	Range
MITDB	024	+0.00221	+0.00222	+0.00220	+0.00221	0.00002
MITDB	031	+0.00342	+0.00506	+0.00447	+0.00432	0.00163
GHL	052	−0.00056	−0.00048	−0.00015	−0.00039	0.00041
SMD	063	−0.05123	−0.04931	−0.05095	−0.05050	0.00192
SVDB	100	+0.08860	+0.08571	+0.08975	+0.08802	0.00404
OPPORTUNITY	135	+0.00092	+0.00045	+0.00053	+0.00064	0.00047
CATSv2	142	−0.01267	−0.01277	−0.01322	−0.01289	0.00055
SMAP	146	+0.00674	+0.00605	+0.00513	+0.00597	0.00160
SMAP	165	+0.00238	+0.00256	+0.00227	+0.00240	0.00029
Exathlon	187	+0.00028	+0.00029	+0.00030	+0.00029	0.00001
Exathlon	193	+0.00085	+0.00085	+0.00083	+0.00085	0.00002

C.4 Five-point strength and bidirectional continuation

We evaluate moment-matched objective weights $\lambda \in \{0, 0.25, 0.50, 0.75, 1\}$. Family-equal changes relative to each sequence's $\lambda = 0$ condition are:

Table C.4: Five-point strength response: family-equal paired VUS-PR changes from each sequence's $\lambda = 0$ condition, with 95% paired bootstrap intervals.

λ	Paired VUS-PR change	95% CI
0.00	0.000	[0.000, 0.000]
0.25	−0.061	[−0.117, −0.016]
0.50	−0.086	[−0.158, −0.021]
0.75	−0.116	[−0.195, −0.041]
1.00	−0.119	[−0.202, −0.043]

All six $\rho(\lambda, V)$ values are negative: -1.0 for MITDB019, LTDB079, LTDB083, and SVDB084; -0.9 for MITDB020 and SVDB085. Four of six five-point estimates are non-increasing and the family-equal five-point estimate decreases strictly.

Paired continuation uses MITDB019, LTDB079, and SVDB084. OO/OM continue the same O20 parent on original/matched data, while MO/MM continue the same M20 parent. Branches reset Adam and share batch order and a 20-epoch budget; complete test sequences are rescored at epochs 0, 5, 10, and 20. Thus $D_O = V_{OO} - V_{OM}$ and $D_M = V_{MO} - V_{MM}$ are exactly zero at epoch 0.

Table C.5: Paired continuation over three sequences and families. D_O and D_M compare branches from identical parents; $V_{MO} - V_{OO}$ measures the remaining restoration gap. Intervals are 95% paired bootstrap intervals.

Epoch	D_O	95% CI	D_M	95% CI	$V_{MO} - V_{OO}$
0	0.000	[0.00000, 0.00000]	0.000	[0.00000, 0.00000]	−0.085
5	+0.061	[0.00752, 0.10585]	+0.010	[0.00094, 0.02371]	−0.074
10	+0.065	[0.01213, 0.11010]	+0.033	[0.01942, 0.05187]	−0.053
20	+0.069	[0.02250, 0.10553]	+0.063	[0.04247, 0.07812]	−0.025

D_O appears quickly and remains, while D_M grows as the model reencounters the original relation. The $V_{MO} - V_{OO}$ gap shrinks but remains negative, supporting that training relations can rewrite score behavior while restoration is partial under the fixed budget.

D PCA-SUBSPACE AND REAL-EVENT VALIDATION

D.1 CONTROLLED SUBSPACE EXPERIMENTS

The experiments reuse trained ShallowAE models and PCA bases for MITDB019, LTDB079, and SVDB084. Each sequence uses five fixed reference samples of 128 windows. The zero-residual condition places queries exactly in the PCA subspace; the constant-residual condition adds the same off-subspace vector to both compared groups.

Table D.1: Score separation between reference and surrogate combinations within a fixed PCA subspace. AUROC first averages reference draws within seeds and then averages the three seed summaries.

Sequence	Residual condition	AUROC
MITDB019	Zero	0.962
MITDB019	Constant	0.941
LTDB079	Zero	0.985
LTDB079	Constant	0.903
SVDB084	Zero	0.994
SVDB084	Constant	0.951

For Figure 4(a), the first two bar groups first aggregate five reference draws within each neural seed and then average the three seed summaries; min–max whiskers span those three summaries. Bar height is the mean minus 0.50 and text inside the bar reports raw AUROC. The whiskers summarize variation across the three seed-level values, with replication defined at the seed summary level rather than the window level.

Maximum coefficient-mean and covariance errors are 1.55×10^{-15} and 6.53×10^{-16}; PCA-score error is 4.26×10^{-14}.

The marginal+correlation-moment construction is generated without access to network scores. Of 15 candidate replicates, 12 satisfy the prespecified covariance- and correlation-moment relative-error threshold of 1%. Qualified-replicate AUROCs are 0.938 for MITDB019, 0.994 for LTDB079, and 0.996 for SVDB084. The third bar group uses qualified seed–construction records: MITDB019 has two qualified constructions and six records; LTDB079 and SVDB084 each have five constructions and 15 records. Means and min–max are computed directly over those records. This construction exactly preserves empirical coefficient marginals and approximately matches covariance and correlation moments, so it is interpreted separately from the first two exact-moment controls. The three remaining MITDB candidates fail the moment-error threshold and are excluded from the summary.

D.2 OBSERVED ANOMALY-EVENT MATCHING

Each contiguous positive-label interval defines one event. Anomaly centers are sampled every 100 timestamps, at most eight per event. Every normal candidate window has all labels equal to zero and centers are not reused. The primary caliper is 0.5% of the normal-prefix PCA-score range $Q_{0.95} - Q_{0.05}$.

Caliper matching constrains the absolute PCA-score difference while allowing small signed imbalances between paired scores. The PCA win rate can therefore differ from 0.50, and the analysis is interpreted as approximate-score matching. ShallowAE win rates first average pairs and neural seeds within events, then weight matchable events equally within families; the family overview finally weights estimable families equally. Win rates and coverage are always reported together.

The primary caliper matches 88/223 events and 129 anomaly–normal centers. At a 0.1% caliper, 52 events remain matchable. The main directions among estimable sequences persist, but LTDB079 coverage falls from 23/116 to 5/116, showing why win rate must be read with coverage.

Table D.2: Complete observed-event matching at the primary caliper.

Family	ID	AE win	PCA win	Difference	Matched/Total	Pairs
Genesis	001	–	–	–	0/3	0
MITDB	019	0.958	0.375	+0.583	1/1	8
MITDB	020	0.900	0.800	+0.100	2/2	10
LTDB	079	1.000	0.391	+0.609	23/116	23
LTDB	083	1.000	0.000	+1.000	3/6	3
SVDB	084	0.817	0.525	+0.292	20/23	27
SVDB	085	0.833	0.426	+0.407	34/57	49
SMAP	144	–	–	–	0/1	0
SMAP	145	0.778	0.333	+0.444	1/1	3
SWaT	171	0.056	0.500	−0.444	2/6	4
Exathlon	174	–	–	–	0/3	0
Exathlon	175	1.000	1.000	0.000	2/4	2

E Relation-Switch Experiments

E.1 Data, coordinates, and paired-training protocol

Each analytical environment uses 2,048 exactly moment-matched training points, 1,024 independent calibration points, and 2,048 independent test points. Both environments share zero centering and one positive scalar training-set RMS c_{rms}: network input is $(u, v)/c_{\mathrm{rms}}$, plotting coordinates and relation distances are returned to raw two-dimensional coordinates, and squared residual scores are computed in standardized coordinates. Writing $x_{\mathrm{raw}} = (u, v)$, the corresponding raw-coordinate map is $r_\theta^{\mathrm{raw}}(x_{\mathrm{raw}}) = c_{\mathrm{rms}} r_\theta(x_{\mathrm{raw}}/c_{\mathrm{rms}})$. Thus

$$\|x_{\mathrm{raw}} - r_\theta^{\mathrm{raw}}(x_{\mathrm{raw}})\|_2^2 = c_{\mathrm{rms}}^2 S_\theta(x_{\mathrm{raw}}/c_{\mathrm{rms}}). \tag{E.1}$$

Relation deviations and approximation bounds scale by c_{rms} between standardized and raw coordinates. The analytical risk $1/3$ is stated in raw coordinates; Equation E.1 gives its relation to standardized scores. The two-dimensional ShallowAE has width 64 and bottleneck dimension 1; each stage trains for 100 epochs. Initial $(+)$ and $(-)$ checkpoints are each continued on both relations to obtain $(++, +-, -+, --)$. Paired branches share parent weights, optimizer-reset rules, and batch order.

E.2 All 18 trained models

Table E.1 groups results by training history, with each model's final training environment as its target. Relation preference is computed on the common fixed union of both test query sets as mean squared distance to the alternative relation minus mean squared distance to the target relation. Positive values mean that reconstructions are closer to the target relation. Distances use constrained analytical projection in raw coordinates.

For the paired comparisons below, both branches are evaluated relative to the same target relation. The retained branch's relation preference therefore changes sign when expressed relative to the switch target. Across the three seeds, switching toward P_+ increases target-relation preference by 0.303, 0.273, and 0.314 relative to staying on P_-. Switching toward P_- gives increases 0.298, 0.264, and 0.289. Both directions simultaneously reverse clear-region score preference from zero to one. Figure 4(c) shows that this reversal develops during continuation across three seeds.

E.3 Score fields, valleys, and calibration boundaries in Figure 4(b)

The main figure shows seed 2026 and two continuations from the same P_+ parent: $P_+ \to P_+$ and $P_+ \to P_-$, both at continuation epoch 100. The shared parent supplies the common initialization for both branches. Both panels share a raw-coordinate grid $u \in [-1.05, 1.05]$, $v \in [-0.62, 0.62]$ with 170 and 130 equally spaced points. Networks receive standardized inputs and directly compute $S_\theta = \|x - r_\theta(x)\|_2^2$. The background displays $\log_{10} \max(S_\theta, 10^{-5})$ over a common range $[-5, 1]$. Threshold and valley calculations use the underlying scores before display clipping.

Table E.1: Relation-switch results across three seeds. The final sign in each training history identifies the target environment. All 18 models achieve AUROC 1.000 in the clear region $|h(U)| \geq 0.15$; full-test AUROC uses the complete test set.

Training history	Seed	Full-test AUROC	Relation preference	Normal FPR
$+$	2026	0.985	0.149	0.045
	2027	0.983	0.122	0.046
	2028	0.985	0.138	0.073
$-$	2026	0.987	0.144	0.040
	2027	0.984	0.131	0.049
	2028	0.984	0.146	0.042
$++$	2026	0.971	0.145	0.048
	2027	0.968	0.128	0.051
	2028	0.978	0.131	0.057
$+-$	2026	0.981	0.154	0.043
	2027	0.979	0.136	0.050
	2028	0.983	0.157	0.052
$--$	2026	0.982	0.140	0.048
	2027	0.982	0.129	0.061
	2028	0.950	0.151	0.039
$-+$	2026	0.983	0.163	0.040
	2027	0.983	0.144	0.061
	2028	0.983	0.163	0.040

For each u column, the actual minimum-score location among the 130 v grid points defines $v^\star_\theta(u)$, without smoothing. We use this discrete minimum only as a finite-grid visualization of the column-wise score valley on the displayed interval. Both panels use identical queries $(-0.75, 0.25)$ and $(0.75, 0.25)$. Black points are queries, white points are reconstructions, and displacements are shown in raw coordinates.

Alarm boundaries use 1,024 independent normal calibration samples from each branch's target environment (P_+ for retention, P_- for switching). Plotting uses the linearly interpolated empirical 95th percentile; FPR evaluation uses the higher empirical quantile.

Each model is calibrated separately, so contour movement combines score-field and threshold changes. Geometric comparisons use the shared-scale score field, valley, and outputs on identical queries; Figure 4(c) supplies threshold-free, bidirectional evidence across seeds.

F Normal Relations and Low-Residual Extension

We compare a bent normal band, $|u| \leq 1$ and $|v - (|u| - 0.5)| \leq 0.03$, with two collinear strips, $0.4 \leq |u| \leq 1$ and $|v| \leq 0.03$, whose central interval is unobserved. Each run uses 2,048 training, 1,024 independent calibration, and 2,048 independent normal test points, centered and scaled by one training RMS. The ReLU AE has hidden width 64 and bottleneck 1, with 100 epochs per training stage and seeds 2026/2027/2028. Each model uses its own normal-calibration Q95.

Figure F.1 contrasts relation-sensitive reconstruction with its limits. In the bent case, the below-threshold region follows the normal relation under original training, changes under moment matching, and returns toward the relation after original-data continuation. In the collinear case, low-score regions can extend across the unobserved gap. Learning a normal relation can therefore reject some unsupported combinations while accepting others along its extension.

We evaluate 16 prespecified cross-gap paths from $(-a, 0)$ to $(a, 0)$, with a evenly spaced from 0.65 to 0.95 and 101 nodes per path. Detection is evaluated at nodes more than 0.06 from the normal support. Original training detects 53.94% of these gap nodes on average, with seed-specific rates of 73.15%, 31.49%, and 57.18%. After matched training followed by original-data continuation (MO),

all three rates are 0%. Paths and nodes serve as quadrature samples, with variation across the three trained seeds reported separately.

Thus a learned low-residual relation can extend into unobserved regions, with the extension depending on initialization and training history. The ideal map $r(u, v) = (u, 0)$ illustrates this boundary analytically in Appendix A.7.

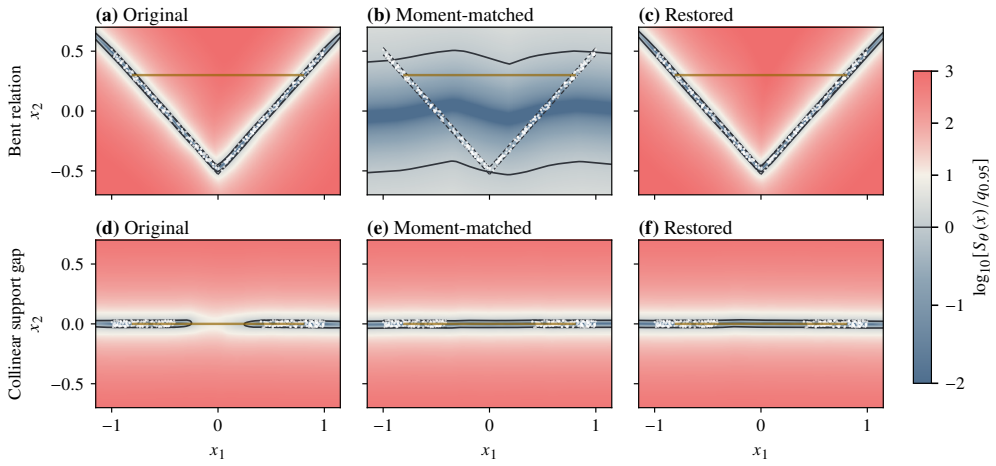


Figure F.1: Normal relations and low-score extensions at seed 2026. Columns show original training (O), moment-matched training (M), and matched training followed by original-data continuation (MO); rows show the bent relation and collinear support gap. Colors display $\log_{10}[S_\theta(x)/q_{0.95}]$, clipped to $[-2, 3]$, using each model's own normal-calibration threshold. Dark solid contours mark $S_\theta = q_{0.95}$; dashed lines delimit normal support, white points show original normal training samples, and gold segments are fixed cross-support queries. Colors compare scores relative to calibration and therefore reflect both the raw score field and the model-specific threshold.