THE SCALE–SHAPE EFFECT ON SLICED-GAUSSIAN REGULARIZER-BASED LATENT PREDICTION METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised regularizers like SIGReg match embedding distributions to isotropic Gaussians by independently penalizing deviations in scale and geometric configuration. However, the substitution rate between these regularization penalties, defined as the scale–shape curvature (τ), is fixed exclusively by the dimensionality of the latent space (called width here). Neither of the theoretical regularizer-objective's kernel (over infinite data and random slice directions), the bandwidth, the slice count, nor the slice dimension can alter this property, and no stationary kernel improves upon the conditioning achieved by the standard baseline. In this work, we propose η, a single weight that rescales the shape component of the objective's kernel, and with it τ, without adding a new penalty term. Consequently, the minimizer remains invariant, the kernel preserves positive definiteness for all $\eta > 0$, and the configuration $\eta = 1$ exactly recovers the standard objective (for e.g. SIGReg). Every η result reported here is obtained on the analytic infinite-slice objective, the closed-form $M \to \infty$ limit of SIGReg (Sec. 2); η has not been applied to the stochastic M-slice estimator itself. Modulating η adjusts the rank of the regularized layer and we show empirical results across two modalities, time-series and image data. This rank modulation is subject to two explicit operational assumptions: sufficient headroom below the rank ceiling, and a non-monotonic layer response, such as an un-regularized backbone preceding a projection head where the regularizer is applied. In downstream evaluations, η shifts performance bi-directionally depending on the specific task and readout mechanism. We therefore present this parameterized axis of variation rather than prescribing an optimal hyperparameter configuration.

1 INTRODUCTION

Regularization methods like SIGReg (Balestriero & LeCun, 2025), map embedding distributions to isotropic Gaussians by penalizing two distinct deviations: incorrect *size*, where the covariance carries an erroneous trace; or an incorrect *shape*, where the covariance carries an erroneous traceless component. However, the substitution rate between these regularization penalties, defined here as the *scale–shape curvature (τ)*, is an intrinsic property of the objective rather than the data. Within the sliced-Gaussian family, this parameter remains fixed and unconfigurable. This ratio is determined strictly by the dimensionality of the latent space (latent width, K). When expanding the population objective (the theoretical regularizer-objective over infinite data and random slice directions) around the target, the ratio of scale curvature to shape curvature evaluates exactly to $\tau_K = (K + 2)/2$ (Sec. 3.1). Consequently, widening a projector from 64 to 1024 dimensions for capacity reasons multiplies the relative penalty of a scale error 16-fold as an unintended side effect.

No existing parameter within such an objective alters this τ (Sec. 3.1), unless for e.g., the distributional target is replaced by a second-moment surrogate. Furthermore, no stationary kernel improves upon this constraint, as the standard objective already achieves the optimal conditioning permitted by its class (Balestriero & LeCun, 2025). Therefore, we introduce a parameter η, that modulates the τ while preserving the target distribution. This parameter reweights the shape component of the objective's reproducing-kernel relative to its orthogonal complement. This formulation acts as an exact decomposition of the baseline objective rather than an additive penalty, ensuring the minimizer remains provably invariant for all $\eta > 0$, with $\eta = 1$ exactly recovering the original regularizer objective.

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.

The practical utility of this axis is currently mixed and is part of our extended future work. It provides a control mechanism over latent geometry, as the rank of the regularized layer responds monotonically across twelve dataset and architecture combinations spanning two input data modalities: time-series and images. Conversely, its impact on downstream accuracy on a given task is inconsistent, shifting performance bi-directionally depending on the task and readout mechanism, or yielding no change at all in some cases. However, an operational parameter that predictably modulates geometry without systematically altering accuracy remains a valuable analytical tool, particularly given the existing evidence on the role of projection heads in regularizer-based methods is predominantly ablative and not fully understood.

Positioning relative to the literature: The most structurally proximate prior research, VISReg (Wu et al., 2026), explicitly reweights scale against shape within a slicing-family regularizer. They conclude that the ideal balance relies inherently on the underlying dataset, observing that weighting towards shape improves performance on degraded datasets while degrading performance on clean datasets. We arrive at a similar conclusion via an alternative construction, thereby reinforcing the fundamental utility of isolating this axis. Our distinct theoretical contribution is the exact mathematical form of the axis and the associated empirical dynamics. Wu et al. (2026) employ an architectural separation, pitting a per-coordinate variance term directly against a sliced-Wasserstein evaluation calculated over stop-gradient-normalized embeddings. Consequently, their scale mechanism penalizes per-coordinate anisotropy, fundamentally deviating from a trace-versus-traceless decomposition. In contrast, our formulation constitutes a rigorous orthogonal decomposition operating entirely within a unified kernel discrepancy. It preserves rotational invariance, guarantees positive semidefiniteness for all $\eta > 0$, mathematically preserves the target law, and exactly recovers the baseline objective at $\eta = 1$ (see Sec. 3). Furthermore, the Wu et al. (2026) sweep configurations while fixing the sum of three discrete coefficients, maintaining a nominal budget but failing to preserve the delivered regularization strength. In our runs the nominal strength is pinned: ρ is calibrated once at $\eta = 1$ and held fixed across the grid, so η is never re-tuned against it. The residual strength confound is then addressed directly by a matched-rank control, in which a pure strength increase that reaches the same rank as the η intervention does not reproduce its downstream effect (App. **??**). Finally, Wu et al. (2026) explicitly omit embedding spectrum measurements, leaving the rank modulation axis functionally unexplored within their framework. Comparisons and expansions to other regularizer-based methods like Weak-SIGReg (Akbar, 2026) e.t.c., is left to future work.

2 SETUP

2.1 THE OBJECTIVE AND ITS DISTINCT COMPONENTS:

To establish the mathematical context, the SIGReg family regularizes an embedding $Z \in \mathbb{R}^{n \times K}$ using a per-slice Epps–Pulley statistic applied to random one-dimensional projections. These projections are resampled at each optimization step and are weighted against the predictive loss by a single regularization strength parameter, ρ (Balestriero & LeCun, 2025) (Epps & Pulley, 1983). To prevent conflation we define the *sliced estimator* as an average computed over M sampled directions. Its *exact infinite-slice limit* is the maximum mean discrepancy.

That limit is an MMD between the batch law P and $G = \mathcal{N}(0, I_K)$ whose kernel is a confluent-hypergeometric function of $\|x - y\|^2$; at the widths used here it is approximated, following Zimmermann et al. (2025), by the inverse-multiquadric kernel

$$k(x,y) = \big(1+\alpha\|x-y\|^2\big)^{-1/2}, \qquad \alpha = \frac{4\gamma}{2K-3},\ \gamma = 1, \qquad D(P) = \mathrm{MMD}_k^2(P,G) = \|\Delta\mu\|_{\mathcal{H}}^2. \tag{1}$$

Numerically, k is evaluated through its exact Gauss–Laguerre representation as a scale mixture of $Q = 48$ Gaussians, $k(x,y) = \sum_q w_q\, e^{-b_q\|x-y\|^2}$ with $b_q = \alpha t_q$ and $\sum_q w_q = 1$; the same mixture gives the functions h and c of Sec. 3.2 in closed form (App. **??**).

Both the exact limit and its approximation are established in the literature (Zimmermann et al., 2025). However, the precise alignment among these components at operational latent widths remains unverified. Crucially, every empirical η result reported in this work is evaluated exclusively on the analytic infinite-slice objective. The η parameter is not applied to the stochastic sliced estimator.

2.2 NOTATION AND THE τ FORMULATION:

The parameter ρ denotes the regularization strength, η represents the proposed shape weight, K defines the latent width. Expressing the perturbed target covariance as $\Sigma = I_K + \epsilon H$, with $H = (\mathrm{tr}H/K)I_K + H_0$ and $\mathrm{tr}H_0 = 0$, the population objective J expands as

$$J = \frac{\epsilon^2}{2}\left[\kappa_{\mathrm{sc}}\frac{(\mathrm{tr}H)^2}{K} + \kappa_{\mathrm{sh}}\|H_0\|_F^2\right] + o(\epsilon^2), \qquad \tau = \kappa_{\mathrm{sc}}/\kappa_{\mathrm{sh}}, \tag{2}$$

where the baseline objective yields $\tau = \tau_K = (K+2)/2$. The τ is defined exclusively at the target $\Sigma = I_K$. We do not read τ directly from trained representations, given that the measured shape curvature can fluctuate in sign, rendering the ratio undefined in practical settings.

2.3 RANK METRIC SPECIFICATION:

Both empirical studies compute $\exp(-\sum_k p_k \log p_k)$ based on an ℓ^1-normalized spectrum p, although the underlying spectra differ by domain. The image evaluations calculate this over the eigenspectrum of the embedding covariance, which is analogous to the standard published rank criterion, albeit substituting singular values with their squares. This yields systematically lower values, and we explicitly refrain from classifying this as an effective rank (Garrido et al., 2023). Conversely, the time-series and trajectory evaluations compute this over the singular-value spectrum of the mean-centered embedding matrix, which aligns directly with the established definition of effective rank (Roy & Vetterli, 2007). These definitions fundamentally represent the same formula applied to different underlying matrices, as explicitly noted in the original rank criterion literature. Importantly, these values are not mathematically interconvertible, as their empirical ratio across our trained representations varies between 1.4 and 8.2. Consequently, we strictly limit our claims to the *response* behavior of η rather than absolute metric levels, normalizing within the respective study whenever both are depicted.

3 THEORY

3.1 τ BLINDNESS

For a fixed latent width $K \geq 2$, consider an objective applying a fixed-target distributional test to unnormalized d-dimensional Haar projections of the embedding. At $\Sigma = I_K$, its τ strictly evaluates to $\tau_K = (K+2)/2$. This value remains independent of the specific per-slice test, the selected kernel and bandwidth, the number of slices in the theoretical regularizer objective, and, provided the kernel is radial, the slice dimension.

Mechanism: Write $\Sigma = I_K + \varepsilon H$ as in Eq. 2. For u uniform on the unit sphere, the projected coordinate $u^\top X$ has law $\mathcal{N}(0,\, 1 + \varepsilon\, u^\top H u)$, so the projected law depends on H only through the scalar $u^\top H u$. Any per-slice discrepancy q that is smooth at the target with $q(1) = q'(1) = 0$ therefore contributes $\frac{\varepsilon^2}{2}\, q''(1)\, \mathbb{E}_u[(u^\top H u)^2] + o(\varepsilon^2)$, and the spherical fourth-moment identity gives

$$\mathbb{E}_u\left[(u^\top H u)^2\right] = \frac{\frac{K+2}{K}(\mathrm{tr}H)^2 + 2\|H_0\|_F^2}{K(K+2)}. \tag{3}$$

Matching Eq. 2 term by term yields $\kappa_{\mathrm{sc}} = q''(1)/K$ and $\kappa_{\mathrm{sh}} = 2q''(1)/(K(K+2))$, hence $\tau_K = (K+2)/2$ immediately, with q entering only through the common factor $q''(1)$. A radial-kernel MMD is the same computation with q replaced by the kernel's radial profile, and a Haar d-dimensional frame V rescales both invariants by the single factor $d(d+2)/(K(K+2))$, which is why neither the test, the kernel, the bandwidth nor d can move the ratio.

A Monte-Carlo probe (see App.A.1) validates both derivations across six distinct latent widths, slice dimensions ranging from 1 to K, and sample sizes spanning a 64-fold range. In the highly discriminating configuration of $K = 1024$ without slicing, the empirical probe recorded $\tau = 513.28$ at $n = 512$ and 512.79 at $n = 2048$, aligning precisely with the theoretical prediction of $\tau = \tau_K = (K+2)/2 = 513.00$ for the chosen K here.

3.2 THE CONSTRUCTION

Let k represent the kernel detailed in Sec. 2, with $\mathcal{H}$ denoting its reproducing-kernel Hilbert space and $\Delta\mu = \mu_P - \mu_G$ denoting the kernel-mean difference. This defines the baseline objective as $D = \|\Delta\mu\|^2_{\mathcal{H}}$. The covariance-shape tangents evaluated at the target G are defined as

$$v_A(y) = y^\top A\, y\, h(\|y\|^2), \qquad A = A^\top,\ \mathrm{tr}A = 0, \qquad \langle v_A, v_B\rangle_{\mathcal{H}} = c\,\langle A, B\rangle_F,\ c > 0, \quad (4)$$

where functions h and constant c are determined in closed form via the same Gaussian mixture that fundamentally defines k. These tangents span a closed, finite-dimensional subspace $\mathcal{H}_{\mathrm{sh}}$. Due to rotational symmetry, the target law holds no component within this subspace, satisfying $\Pi_{\mathrm{sh}}\mu_G = 0$. By decomposing $\Delta\mu$ along the $\mathcal{H}_{\mathrm{sh}}$ subspace and reweighting the corresponding projection, we define

$$D_\eta = \|(I - \Pi_{\mathrm{sh}})\Delta\mu\|^2_{\mathcal{H}} + \eta\,\|\Pi_{\mathrm{sh}}\Delta\mu\|^2_{\mathcal{H}}, \qquad \eta > 0. \quad (5)$$

This establishes a strict orthogonal decomposition of the existing objective rather than appending a supplementary penalty. The modified objective D_η functions fundamentally as a squared maximum mean discrepancy under the modified kernel $k_\eta = k_\perp + \eta\, k_{\mathrm{sh}}$, which is derived directly by reweighting the projection of the underlying feature map. Consequently, it guarantees positive definiteness for all $\eta > 0$. It maintains characteriztic behavior, as established by the bounds

$$\min\{1,\eta\}\, D \;\le\; D_\eta \;\le\; \max\{1,\eta\}\, D, \qquad \text{so} \qquad D_\eta = 0 \iff P = G, \quad (6)$$

which strictly preserves the target distribution. Evaluating at $\eta = 1$ recovers the baseline formulation D, positioning the standard objective as one case on a continuous axis rather than an isolated baseline. At the target $\Sigma = I_K$, the scale curvature is maintained perfectly while the shape curvature scales directly by η, yielding

$$\tau \;=\; \frac{\tau_K}{\eta} \;=\; \frac{K+2}{2\eta}. \quad (7)$$

Reweighting a kernel discrepancy via a positive, bounded operator acting on its reproducing-kernel Hilbert space is a known mathematical mechanism (Hagrass et al., 2025), and target preservation inherently follows for any operator bounded rigorously above and below, as demonstrated by Eq.(6). We claim neither the basic reweighting mechanism nor the preservation of the target. Our specific contributions encompasses the specific curvature definition in Eq.(7), which grounds η specifically as a τ parameter rather than an arbitrary weighting coefficient. While Eq.(7) is formulated precisely at $\Sigma = I_K$, empirical evaluations indicate that measured shape curvature on trained representations can invert in sign. Thus, we do not read τ directly from η in practice. Instead, the parameter's practical utility relies strictly on the rank response described in Sec. 4.1.

4 EXPERIMENTS AND RESULTS

The empirical studies were performed on four existing time-series datasets, `ETTh1` (Zhou et al., 2021a) and the `C-MAPSS` (Frederick et al., 2007) subsets `FD001`, `FD002`, and `FD003`; as well as the `CIFAR` (Krizhevsky et al., 2009) and `MNIST` (LeCun & Cortes, 2010) image datasets. All results are reported at five initialization seeds each across two waves comprising 160 runs. We additionally evaluated this on thirteen distinct trajectory tasks, encompassing two modalities and four architectures.

The sweep-space was defined as the paired within-seed span $\mathrm{rank}(\eta{=}30) - \mathrm{rank}(\eta{=}0.1)$ measured at the conclusion of the 100-epoch schedule, maintaining fixed parameters: latent width, $K = 64$ and the regularization strength, $\rho = 0.1$. Because ρ was strictly pinned across the parameter grid within each evaluation cell by design, this rank modulation cannot be dismissed as an obfuscated strength effect yet and this study is part of our future work.

4.1 THE RANK RESPONSE

Validating whether η reliably modulates the target geometry is the primary goal across both input modalities here. This functions strictly as an operational manipulation check on the parameter, rather than a downstream empirical result. The rank metrics presented for time-series and image data is as described in Sec. 2.3. Across our evaluated trained representations, the ratio between these two definitions varies from 1.4 to 8.2. Figure 1 graphs each dataset's rank (a) and task (b) response relative to its own baseline at $\eta = 1$ (SIGReg), with absolute values cataloged in Tab. 1.

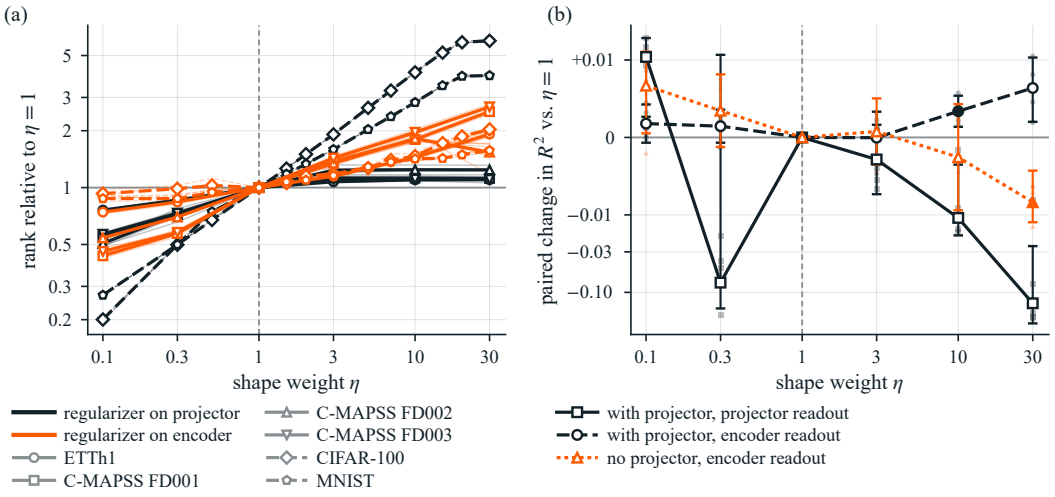

Figure 1: The parameter η governs the horizontal axis for both logarithmic panels, with the default configuration (SIGReg) explicitly marked. Panel (a) displays the rank of the layer targeted by the regularizer across twelve dataset and architecture cells over two modalities. The anomalous turnover series (between $\eta = 10$ and 30) corresponds to `C-MAPSS FD002` lacking a projector. Panel (b) depicts the downstream performance response on `C-MAPSS FD001`, evaluated as a paired contrast against η=1 across 5 seeds. Measurements capture readouts at the projector and the corresponding encoder for models with projectors; and the encoder for models without projectors.

4.1.1 HEADROOM REQUIREMENT:

The η parameter successfully modulates rank only when sufficient headroom remains below the strict ceiling defined by the latent width. It fails to induce movement if the regularizer has already pinned the rank against this ceiling. As seen in Fig. 2, at $K = 4$, the empirical rank fluctuates minimally between 3.93 and 4.00 across the full parameter grid over all evaluated tasks. By contrast, at $K = 64$, the rank reliably shifts by 12 to 36 units across all tasks. The `FD001` dataset with a designated projector exhibits this same saturation limit, plateauing at 63.06 out of a strict maximum of 64.

4.1.2 TARGET LAYER SPECIFICITY:

The geometric response maintains monotonicity solely within the specific layer the regularizer directly acts upon. When evaluated on the un-regularized backbone situated behind the projector, the identical models exhibit non-monotonic behavior, peaking in rank near $\eta = 3$ before declining by 16 to 31 units. This distinct interior peak replicates across two trajectory tasks and the image evaluations, representing three independent sources across two distinct modalities. On the first C-MAPSS subset without a projector, on the conventional asymmetric engine score, the family's default is the worst of the seven arms: ranked best to worst, the unregularized anchor 1507.7, then $\eta = 0.1$ at 1511.4, $\eta = 10$ at 1546.9, $\eta = 30$ at 1561.9, $\eta = 0.3$ at 1685.6, $\eta = 3$ at 1698.5, and $\eta = 1$ at 1965.7. Seed by seed the default ranks fifth, sixth, fourth, sixth and sixth of six. A rank-based omnibus test gives $p = 0.0160$, and the multiplicity-safe statement that the default is worse than the best other arm gives $p = 0.0051$. The best pairwise contrast is $\eta = 0.1$ at $+454.29$, 95% interval $[+7.89, +900.69]$, nominal $p = 0.0476$ against a corrected threshold of 0.01.

In the `FD002` dataset configuration lacking a projector, the rank measurement increases up to $\eta = 10$ but falls at $\eta = 30$ consistently across $5/5$ seeds, peaking at 15.7 % of the true latent width, explicitly disproving rank saturation as the cause. Given this ambiguity, we plot this curve explicitly in Figure 1(a) without smoothing and abstain from asserting universal monotonicity across all cells. Resolving this anomaly definitively requires additional seed runs (left for future work).

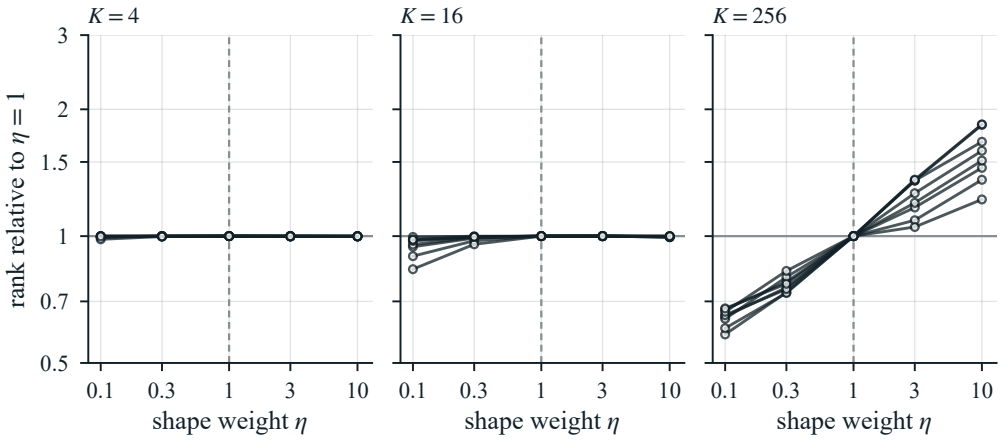

Figure 2: Rank of the regularized layer against the shape weight, for a list of 8 dynamical trajectory datasets (D4RL, DeepMind Control (Fu et al., 2021) (Tassa et al., 2018)) and toy attractors as an initial exploratory analysis at three projector widths, three seeds per point, strength $\rho = 0.1$ held fixed across the grid. At the widest setting the knob moves rank by a factor of three; at a sixteenth of that width the widest span across the entire grid is 1.7% of the widest setting, because the regularizer has already pinned the spectrum and there is nothing left to move.

Table 1: Downstream paired contrasts aggregating both time-series studies under uniform multiplicity correction.

Data	Config.	Readout (layer)	Tier	η	Δ vs η=1	95 % CI	99 % CI	seeds	floor
ETTh1	proj.	96-step rollout err. (projector)	pre-reg.	30.0	+0.1258	[+0.0335, +0.2181]	[-0.0272, +0.2788]	5/5	±0.0261
ETTh1	proj.	96-step rollout err. (projector)	pre-reg.	0.1†	+0.1032	[+0.0620, +0.1444]	[+0.0348, +0.1716]	5/5	±0.0261
ETTh1	no proj.	96-step rollout err. (encoder)	pre-reg.	3.0	-0.0503	[-0.1902, +0.0896]	[-0.2822, +0.1816]	3/5	±0.0785
FD001	proj.	RUL var.-explained (projector)	pre-reg.	0.1	+0.0109	[+0.0027, +0.0192]	[-0.0028, +0.0246]	5/5	±0.0052
FD001	no proj.	RUL var.-explained (encoder)	pre-reg.	0.1	+0.0067	[+0.0005, +0.0128]	[-0.0035, +0.0169]	4/5	±0.0027
FD003	proj.	RUL var.-explained (projector)	pre-reg.	3.0	-0.0071	[-0.0196, +0.0054]	[-0.0278, +0.0136]	4/5	±0.0047
FD003	no proj.	RUL var.-explained (encoder)	pre-reg.	0.3	-0.0013	[-0.0040, +0.0014]	[-0.0058, +0.0032]	4/5	±0.0021
FD002	proj.	RUL var.-explained (projector)	pre-reg.	3.0	-0.0049	[-0.0178, +0.0081]	[-0.0263, +0.0166]	3/5	±0.0046
FD002	no proj.	RUL var.-explained (encoder)	pre-reg.	3.0	+0.0044	[+0.0004, +0.0085]	[-0.0023, +0.0112]	4/5	±0.0039
FD001	proj.	RUL RMSE (cycles) (encoder)	secondary	30.0	+0.3197	[+0.1012, +0.5382]	[-0.0426, +0.6821]	5/5	±0.1291
FD001	proj.	RUL RMSE (cycles) (encoder)	secondary	10.0†	+0.1688	[+0.0689, +0.2688]	[+0.0031, +0.3346]	5/5	±0.1291
FD001	proj.	RUL RMSE (cycles) (projector)	secondary	0.1	+0.5434	[+0.1321, +0.9546]	[-0.1387, +1.2254]	5/5	±0.2562
FD003	proj.	RUL RMSE (cycles) (encoder)	secondary	0.1	+0.4015	[+0.0870, +0.7160]	[-0.1200, +0.9230]	5/5	±0.2287
FD003	proj.	RUL RMSE (cycles) (encoder)	secondary	30.0†	+0.3255	[+0.2228, +0.4283]	[+0.1551, +0.4960]	5/5	±0.2287
FD002	proj.	RUL RMSE (cycles) (encoder)	secondary	0.3	+0.1028	[+0.0285, +0.1772]	[-0.0205, +0.2261]	5/5	±0.0683
ETTh1	proj.	rollout-mean err. (projector)	secondary	0.1	+0.0827	[+0.0495, +0.1160]	[+0.0277, +0.1378]	5/5	±0.0220
ETTh1	no proj.	reconstruction err. (encoder)	secondary	30.0	+0.0193	[+0.0070, +0.0315]	[-0.0010, +0.0395]	5/5	±0.0067
ETTh1	no proj.	reconstruction err. (encoder)	secondary	10.0†	+0.0113	[+0.0050, +0.0176]	[+0.0009, +0.0217]	5/5	±0.0067

4.2 Downstream task results for time-series data

Modulating η impacts downstream performance inconsistently. The observed shifts occur in both directions depending heavily upon the specified task and readout mechanism, while numerous configurations show zero statistical movement. On time-series data, the final tally yields zero positive outcomes even in cases where the rank is monotonic (Fig. 1(b)). The uncorrected tally resolves to four out of eight (Table 1). On CIFAR data, in the few-shot regime there is a marked rise in performance over SIGReg ($\eta = 1$), and performance falls again beyond this region as seen in Fig. 3(b), while the rank increases monotonically with η (Fig. 3(a)).

4.2.1 Analysis of statistically resolved movements:

The clearest statistically resolved performance shifts occur in secondary readouts rather than the primary layer outputs. For instance, on the FD001 dataset with η affecting the designated projector, the backbone encoder-read error metric improves at the top of the parameter grid by $+0.32$

$[+0.10, +0.54]$ cycles across all 5 seeds relative to SIGReg. Compared to the unregularized baseline, this improvement expands to $+0.98$ $[+0.59, +1.36]$. Importantly, this phenomenon replicates on a separate C-MAPSS subset, where the same encoder readout yields an improvement of $+0.33$ $[+0.22, +0.43]$ across all 5 seeds. See Table 1.

We also identify anomalous results on C-MAPSS subsets where performance does not improve on the projector, but instead on the backbone encoder, while rank is still monotonic. Therefore, while modulating η does influence downstream behavior, it cannot operate through geometric rank modification alone (Garrido et al., 2023; Jing et al., 2022), and the performance modulation differs based on where the readout probe is on a case-by-case basis. See App. B for further results and discussions on this.

4.3 Image modality results

We show results on the backbone in a 10-shot regime, on held-out test queries, $\eta = 2$ against $\eta = 1$ at $\rho = 7.5$ on an 80-epoch schedule, paired within seed. All results are shown using residual networks (He et al., 2016) on CIFAR (Krizhevsky et al., 2009) and MNIST (LeCun & Cortes, 2010) and are meant to provide empirical support. Expanding to larger benchmarks is left to future work.

Table 2: All three image-results are five seeds; all fifteen seeds are positive.

setting	phase	paired change	note
CIFAR-100 / ResNet-18	Q	$+0.80$ $[+0.67, +0.93]$	5/5 seeds; whitened $+0.81$
CIFAR-10 / ResNet-18	R	$+0.41$ $[+0.25, +0.58]$	5/5 seeds; below the committed $+0.5$ rule
CIFAR-100 / ResNet-34	S	$+0.85$ $[+0.59, +1.10]$	5/5 seeds; whitened $+0.77$
random-effects pool	Q,R,S	$+0.68$ $[+0.41, +0.96]$	$I^2 = 93\%$ — the three means are the result, not the pool
full-data logistic probe	Q,R,S	$+0.44$ $[+0.35, +0.54]$	$I^2 = 0\%$; components $+0.39/+0.56/+0.41$

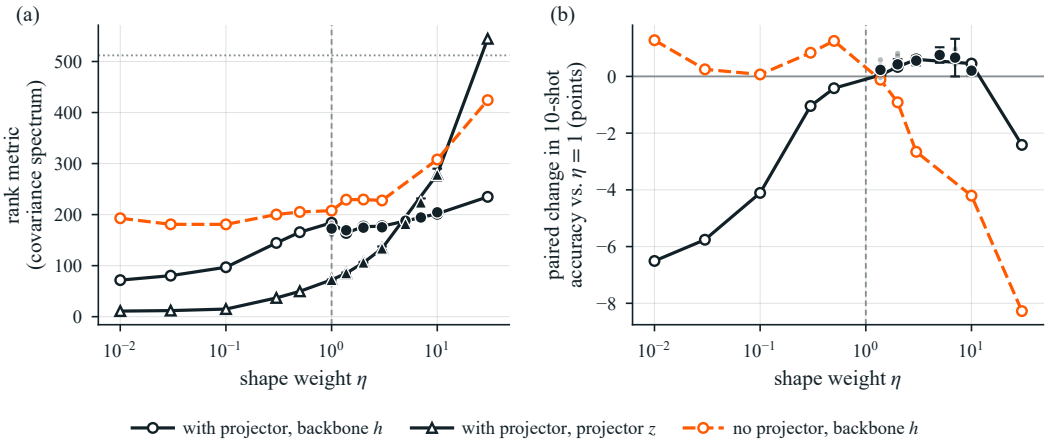

Figure 3: ResNet-34 on CIFAR-100, $\rho = 7.5$, 320 epochs. η is the horizontal axis of both panels, logarithmic, with the default (SIGReg) marked. **(a)** The RankMe metric. **(b)** A 10-shot ridge on backbone features evaluated on held-out test queries, as a paired difference from the $\eta = 1$ (SIGReg) from the same 5 seed runs shown in panel (a). Two effects of one intervention; neither explains the other, and we do not claim the change in rank produces the change in accuracy. Open markers joined by a line are a single seed spanning the whole η range; filled markers with bars are means with 95% t intervals over 5 seeds. The display is truncated at $\eta = 30$, the largest value at which gradient clipping stays under 1% of steps in both conditions.

At $\rho = 15$ over 80 epochs the projector-side rank metric reads 57.3, 82.9 and 123.1 at $\eta = 0.5$, 1 and 2, and the backbone 53.6, 58.9 and 67.0; the paired projector change from $\eta = 1$ to $\eta = 2$ is $+40.2$

$[+39.5, +40.9]$ on 5/5 seeds, and from $\eta = 1$ to $\eta = 0.5$ it is -25.6 $[-26.6, -24.6]$ on 0/5; η moves rank in *both* directions from the default.

As mentioned in Sec. 2.3, the image results report the exponentiated Shannon entropy of the eigen spectrum of the embedding covariance: the published rank criterion with the singular values replaced by their squares, and systematically lower than it (Garrido et al., 2023). We do not call it effective rank.

Figures 3 gives the response for the 10-shot accuracy on CIFAR-100.

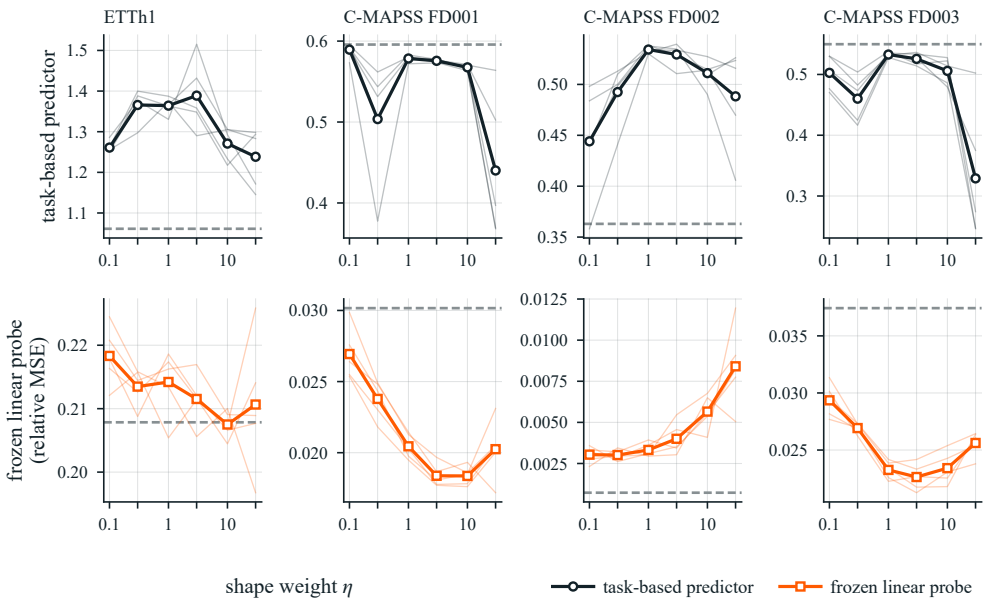

Figure 4: The task based predictor values (top): 96-step forecast-based relative MSE for ETTh1 and RUL-regression for the 3 C-MAPSS subsets; and a frozen-representation linear probe (bottom) on the same trained models, four datasets, with-projector configuration, five seeds drawn individually, 100 epochs. The dashed line is the unregularized anchor. On three of the four datasets the anchor beats every regularized variant when evaluated from the task=specific predictor; on two of those, the frozen probe reverses observation, and on the fourth both readouts point in opposite ways. Vertical axes are absolute and per panel because the two readouts are in different units.

5 Discussions

We propose an instrument η for sliced-Gaussian regularizer based methods that demonstrably adjusts the defined rank metric wherever the network architecture affords sufficient operational capacity. It correspondingly shifts downstream evaluations across a subset of specific tasks and readouts, moving in task-dependent directions while remaining neutral in others. Currently, no measurement robustly predicts the parameter's impact on performance on a novel task or dataset, but we empirically show the expected rank metric trend on it.

5.1 Operational utility of the parameter:

This intervention serves two distinct operational purposes. First, researchers can sweep the η axis empirically. Because the axis is strictly one-dimensional and mathematically bounded, and because $\eta = 1$ precisely recovers the baseline objective (SIGReg in our case), sweeping the parameter provides an inherent control arm. Second, we position the construction as an analytical instrument rather than a validated phenomenon. Because η controls geometric rank reliably without systematically compromising accuracy, it allows researchers to investigate complex mechanistic hypotheses, such as identifying the optimal layer for regularization or probing the explicit function of the projection

head, across a continuous spectrum rather than relying exclusively on structural ablations. While our naive test on layer-specific regularizer preferences failed to generalize, the mathematical tool remains fundamentally sound.

5.2 LIMITATIONS:

The chosen readout probe decides both whether regularizing helps and which way to tune η as shown in Fig.4; beyond that, the rank response needs headroom below the ceiling and leaves one cell unresolved in both directions, and every empirical result is shown at fixed small scales: 32×32 images and small sequence models.

First, the strictly unregularized anchor matched or exceeded the performance of all regularized arms in three out of the eight evaluated cells, all of which utilized a projector. Furthermore, it demonstrated statistical resolution in two of these three scenarios. Conversely, the regularized models proved universally equivalent or superior across all four projector-free configurations. Evaluated via a frozen linear probe on the exact same trained models, this conclusion reversed drastically. The standard regularized default outperformed the unregularized baseline across all eight cells, achieving clear statistical resolution in seven (Appendix C). We report both empirical realities without enforcing a selection hierarchy.

Second, the preferential selection of the evaluation readout governs both the presence of statistical movement and the ideal directionality of the η parameter. Regarding the evaluated forecasting dataset, only two of the four defined rollout readouts register any statistical movement, and these two metrics strictly prefer opposite extremums of the parameter grid.

Third, our evaluations operate strictly within defined scaling limits. Image evaluations utilize 32×32 inputs across schedules ranging from 80 to 320 epochs, while time-series benchmarks rely on constrained architectures operating over 100 epochs. All derived measurements are extracted precisely at the conclusion of fixed schedules rather than at theoretical convergence limits.

Fourth, the rank modulation remains strictly dependent on architectural headroom. Operating at a width of $K = 4$, the maximum span across the entire η parameter grid equals a marginal $1.7\,\%$ of the theoretical ceiling, as the baseline regularizer rigidly dictates the spectrum and pre-empts further modulation.

Fifth, one isolated rank evaluation cell lacks definitive resolution. On the second `C-MAPSS` subset configured without a projector, the rank measurement peaks explicitly at $\eta = 10$ and declines by $\eta = 30$ across all 5 initialization seeds. This decline occurs at $15.7\,\%$ of the theoretical ceiling, explicitly ruling out saturation. However, the calculated paired contrast evaluates to -12.27 with a broad $95\,\%$ confidence interval of $[-28.24, +3.71]$, rendering the decline statistically indistinguishable from zero or an increase. We explicitly report this discrepancy without claiming it as a definitive failure or success.

5.3 FUTURE WORK:

Three questions follow directly from our results here. First, nothing we measured predicts which way η will move accuracy on an unseen task; testing the rank-band reading above needs tasks whose suitable rank is known independently of the outcome being predicted, which none of our settings provides. Second, the dial currently requires the population objective; above $d = 1$ the projected defect carries two independent quadratic invariants, so a sliced form should exist, and making it target-preserving is open. Third, the cells we leave unresolved, the turnover at the top of the grid on one `C-MAPSS` subset and whether the layer a representation is read at has a preference of its own, call for separate experiments with identical-configurations and enough seeds, rather than seeds added to the studies reported here. The scale of the present evidence is also modest: 32×32 images and small time-series models, read at the end of fixed schedules. Addressing these shortcomings is part of our future work.

AI USE STATEMENT

In this work agentic coding tools to draft and automate the code based experiments. The code was verified using human-authored code harnesses (as a unit-test style quality control). The results

generated were evaluated and validated by human authors. Generative AI was used to draft the initial skeleton of this manuscript with human authors responsible for major line-by-line revisions of the text and plots and the other authors provided editorial passes on the human-revised version, blind to the initial AI generated draft. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Here are the tasks we used AI assistance for:

Generate synthetic data sets (only for a minor exploratory result quoted), formulate mathematical claims, provide critical ingredients for proving mathematical claims (to formalize and verify setup and theory sections), assist in the writing of proofs implement methods (theory), support qualitative and thematic data analysis (plot generation via code), create or modify scientific figures or images (via code), create or edit software code, draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, propose keywords for a research paper.

Reproducibility statement

The anonymized code archive: `https://anonymous.4open.science/r/wtreg-2246/`, contains the code and the self generated trajectories for reproducibility, the other datasets used are publicly available.

References

Habibullah Akbar. Weak-SIGReg: Covariance Regularization for Stable Deep Learning, March 2026. URL `http://arxiv.org/abs/2603.05924`. arXiv:2603.05924 [cs.LG].

Randall Balestriero and Yann LeCun. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics, November 2025. URL `http://arxiv.org/abs/2511.08544`. arXiv:2511.08544 [cs.LG].

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, January 2022. URL `http://arxiv.org/abs/2105.04906`. arXiv:2105.04906 [cs.CV].

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.3.723. URL `https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.3.723`.

Dean K Frederick, Jonathan A DeCastro, and Jonathan S Litt. User's guide for the commercial modular aero-propulsion system simulation (c-mapss). Technical report, 2007.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for Deep Data-Driven Reinforcement Learning, February 2021. URL `http://arxiv.org/abs/2004.07219`. arXiv:2004.07219 [cs.LG].

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank, June 2023. URL `http://arxiv.org/abs/2210.02885`. arXiv:2210.02885 [cs.LG].

Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li. Spectral Regularized Kernel Goodness-of-Fit Tests, January 2025. URL `http://arxiv.org/abs/2308.04561`. arXiv:2308.04561 [math.ST].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL `http://ieeexplore.ieee.org/document/7780459/`.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning, April 2022. URL `http://arxiv.org/abs/2110.09348`. arXiv:2110.09348 [cs.CV].

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann</inline> LeCun and Corinna</inline> Cortes. Mnist handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

Olivier Roy and M. Vetterli. The Effective Rank: A Measure Of Effective Dimensionality. September 2007. doi: 10.5281/ZENODO.40328. URL http://zenodo.org/record/40328.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite, January 2018. URL http://arxiv.org/abs/1801.00690. arXiv:1801.00690 [cs.AI].

Haiyu Wu, Randall Balestriero, and Morgan Levine. VISReg: Variance-Invariance-Sketching Regularization for JEPA training, June 2026. URL http://arxiv.org/abs/2606.02572. arXiv:2606.02572 [cs.CV].

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021a.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, May 2021b. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. KerJEPA: Kernel Discrepancies for Euclidean Self-Supervised Learning, December 2025. URL http://arxiv.org/abs/2512.19605. arXiv:2512.19605 [cs.LG].

A Empirical validation of the theory

A.1 Monte-Carlo agreement.

A population probe, no training, over six latent widths, slice dimensions spanning 1 to K and sample sizes over a 64-fold range agrees with both derivations. The discriminating cells are the no-slicing endpoints:

Table 3: Measured τ of the kernel family at $d = K$, against the predicted $(K + 2)/2$. The second-moment statistic evaluated at the same cells returns 1.000.

K	d	n	measured	predicted
1024	1024	512	513.283	513.0
1024	1024	2048	512.792	513.0
512	512	512	257.964	257.0
512	512	2048	257.124	257.0
256	256	512	128.951	129.0
256	256	2048	128.976	129.0

The $d = 1$ column of the same sweep runs a few percent high at every width and in both statistic families, and the excess grows with K (from $+1.1$ to $+2.4\%$ at $K = 16$, up to $+47\%$ at $K = 1024$, $n = 512$). This is a property of the probe, not of the statistic: it is a ratio of two noisy second differences whose denominator is τ_K times smaller than the numerator, so the estimator is biased upward exactly where the denominator is smallest. A step-size sweep over a factor of eight shows no trend, and the excess is the same size in both families at every width.

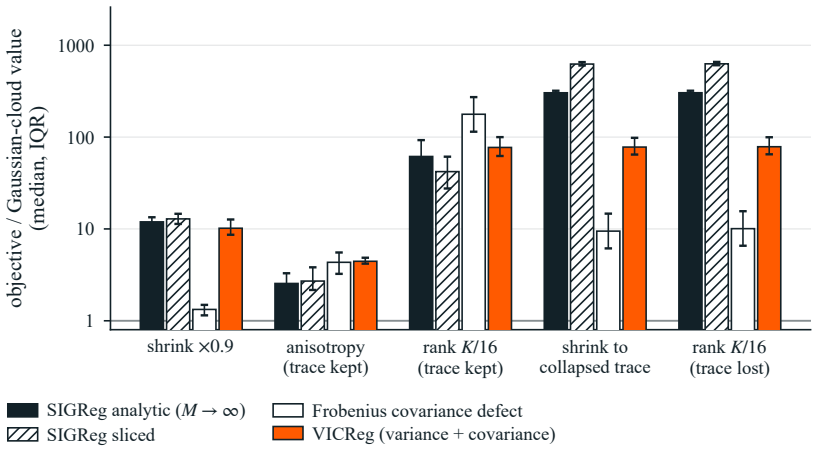

Figure 5: Median and interquartile range of four objectives under five controlled perturbations of a whitened embedding cloud, as a multiple of the objective's value on an exactly Gaussian cloud, over the same 417 trained runs. The variance-plus-covariance term of the second-moment family returns the same number for a rank collapse with the trace preserved, a rank collapse with the trace destroyed, and a pure shrinkage to that same trace (77.1, 78.5 and 77.8); the sliced-Gaussian objectives separate them by an order of magnitude (61.5 against 304 and 304). Logarithmic vertical axis. (Bardes et al., 2022)

B Downstream task results

See Tab. 1. Evaluation contrasts represent paired within-seed measurements anchored against η=1. Values are signed such that positive figures denote improvement over the default baseline. Confidence intervals represent two-sided paired $t(4)$ distributions, while the 99 % column reflects the multiplicity correction applied uniformly over the five non-default grid parameters. The † symbol designates configurations surviving correction that do not correspond to the absolute maximal mean, presenting both sets to prevent any perception of post-hoc selection based on resolution outcomes. The secondary block is explicitly excluded from all formal tallies, representing readouts designated subsequent to initial data capture. The paired `ETTh1` committed rows reflect a designated readout where all models, including the unregularized baseline, fail to surpass a simple mean predictor. Rows lacking a projector sample a 512-dimensional layer, while projector rows sample a 64-dimensional layer, rendering direct architectural comparisons invalid at fixed latent widths.

B.1 Quantifying the magnitude of the shape intervention:

To precisely contextualize the influence of the η parameter, we evaluate the cell demonstrating the clearest positive resolution. Opting to deploy the regularizer at its default setting yields an improvement of 0.66 cycles $[+0.28, +1.04]$. Exploiting the entire available η grid affords a marginal additional improvement of 0.32 cycles. Relative to the baseline metric level of 40.90 cycles, these improvements constitute 1.6 % and 0.8 %, respectively. These distinct gains sum perfectly to the $+0.98$ total cited previously. Crucially, the baseline improvement is unstable regarding its sign. Across three of the five other comparable `C-MAPSS` cells utilizing this specific readout, the direction inverts, with the strictly unregularized anchor demonstrating superior performance.

B.2 The counterweight evaluation perspective:

The completely unregularized anchor matches or outperforms all regularized models in three out of the eight core dataset-by-configuration cells. Notably, all three of these instances incorporate a projector, with two achieving statistical resolution. Conversely, the regularized models strictly match or outperform the unregularized baseline across all four configuration cells lacking projectors. However, when assessing these identical trained models via a frozen linear probe, this outcome definitively reverses. Under this probe evaluation, the default regularized setting outperforms the

unregularized anchor across all eight evaluation cells, demonstrating clear resolution in seven out of the eight, frequently by factors spanning three to seventy. Both evaluation metrics remain empirically valid. The selected evaluation methodology dictates the resulting conclusion, underscoring our broader evaluation thesis applied to a distinct modality. We explicitly position this duality as the primary limitation of our analysis.

B.3 Disconfirming rank as the primary performance mediator:

A controlled variation of regularization strength induces a strictly larger rank alteration across both measured layers than the maximum displacement achievable via the η parameter. The resulting ranks match within $1.6\,\%$ on the backbone and $4.1\,\%$ on the projector, with the residual gap favouring the strength control arm. Despite generating higher rank, this strength intervention yielded inferior results, underperforming the shape arm by $+0.74$ $[+0.64, +0.85]$ across $5/5$ seeds. This specific disparity replicated on a secondary image dataset. Consequently, a backbone-rank increase of $+7.06$ generated a marginal $+0.06$ improvement, whereas a smaller rank increase of $+6.15$ achieved a $+0.80$ improvement. Furthermore, independent analysis reveals that elevated rank degrades low-shot usability until the representations undergo explicit conditioning correction. Specifically, raw higher-rank projector features underperform baseline features given ≤ 20 labels, only achieving superior performance post-whitening.

B.4 An observation failing subsequent replication

Evaluating the `FD001` dataset, identical models employing identical initialization seeds and predictions exhibit diametrically opposed parameter preferences solely contingent upon the evaluated layer. The projector-layer readout optimizes at $\eta = 0.1$, recording an improvement of $+0.0109$ across $5/5$ seeds, representing a $2.1\times$ floor threshold. Conversely, the backbone-layer readout optimizes at $\eta = 30$, registering $+0.0064$ across $5/5$ seeds, marking a $2.5\times$ threshold. While both results exhibit uncorrected significance, neither survives the formal multiplicity correction enforced in Table 1. We attempted to replicate this specific comparison across two of the three remaining time-series datasets. The fourth dataset fundamentally lacks a target channel, rendering the requisite readout family mathematically uncomputable. On the two viable datasets, the layer divergence completely fails to reproduce. No configuration of $\eta \neq 1$ improves the corresponding projector readout under any tested label budget. Furthermore, the backbone readout on both datasets optimizes at values explicitly below the baseline default. Thus, the apparent directional split devolves into a simple disparity between an observable preference and the complete absence of a preference. Because the backbone-side readout variance on these subsequent datasets matches or exceeds the variance observed on `FD001`, we definitively categorize this as a non-replication rather than a methodological limitation incapable of capturing the effect.

Standard architectural analyses identify dimensional collapse, map it spectrally, and verify that the projector prevents this collapse from propagating into the core representation (Jing et al., 2022). The accepted remedy actively removes the trainable projector, achieving superior baseline performance, proving this is not strictly an ablative phenomenon. The η parameter theoretically introduces a continuous gradient of control at a fixed latent width, contrasting with discrete structural modifications. Based on our current empirical evidence, this continuous parameter facilitates new avenues for interrogating network dynamics, though it does not yet provide definitive answers.

C Additional results

The figures below support claims made above and in the main text. They are placed together so that the lettered appendices keep the structure the main text refers to.

C.1 Per-dataset time-series curves

All four are built to one design so that they can be compared: the rank metric of the read layer against η, and the downstream readouts against η on the same horizontal axis, with five individual seeds drawn behind each mean and the projector width ruled. The time-series-side rank metric is the *effective rank*, the exponentiated Shannon entropy of the normalized singular-value spectrum, of

the mean-centered embedding matrix, which is the published rank criterion's formula applied to the centered matrix (Roy & Vetterli, 2007) (Garrido et al., 2023).

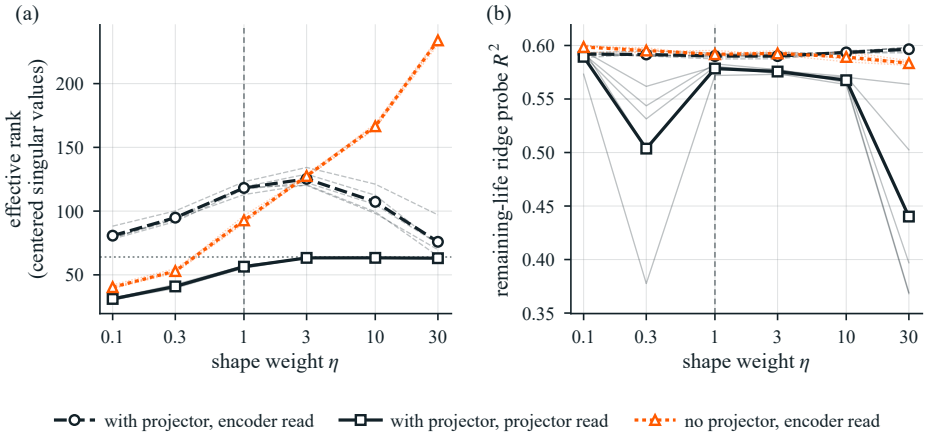

Figure 6: (a) Effective rank of each read layer against the shape weight; the dotted line is the projector width $K = 64$, and the encoder width of 512 is off the top of the axis. (b) The downstream readouts on the same horizontal axis. Five seeds drawn individually, 100 epochs, strength $\rho = 0.1$ held fixed across the grid by construction. Neither panel explains the other.

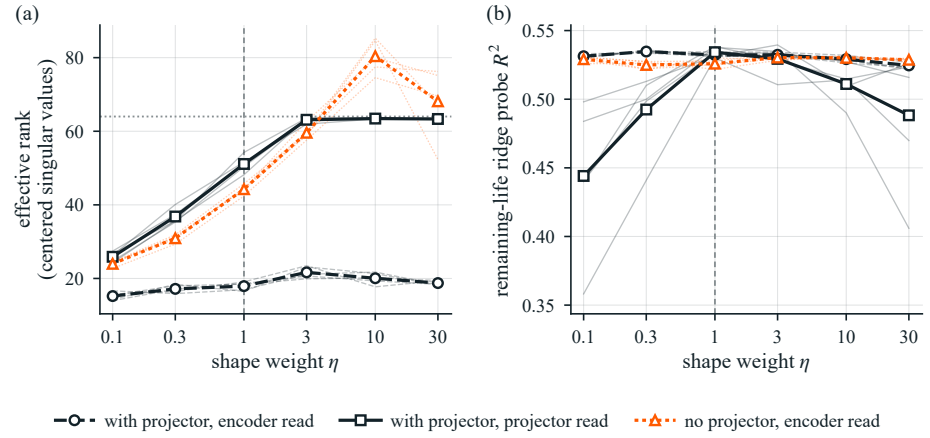

Figure 7: Panels, seeds, schedule and strength as in Figure 6. Without a projector the rank response turns over at the top of the range; the fall is not distinguishable from zero, nor from an increase.

C.2 The layer disagreement does not replicate

On the first C-MAPSS subset, reading the same trained models at the regularized projector and at the deployed backbone gives opposite preferences over η, best at 0.1 and at 30 respectively, $+0.0109$ and $+0.0064$, both 5/5 seeds, 2.10 and 2.47 times their measured floors, uncorrected. The comparison could be repeated on two of the three remaining time-series datasets, the fourth carries no remaining-life readout at all and cannot test it in either direction. On neither of the two does the split appear: no $\eta \neq 1$ improves the projector readout at all on either dataset, at any of three label budgets, while the backbone readout prefers a setting *below* the default on both. The readout ranges on the new datasets are at least as large as on the original, the backbone-side span, in units of that cell's floor, is 6.06 and 2.07 against 2.48.

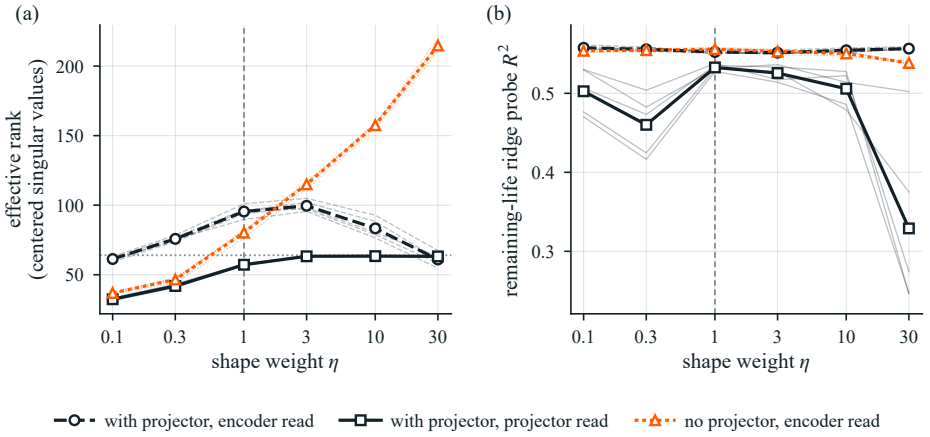

Figure 8: Panels, seeds, schedule and strength as in Figure 6.

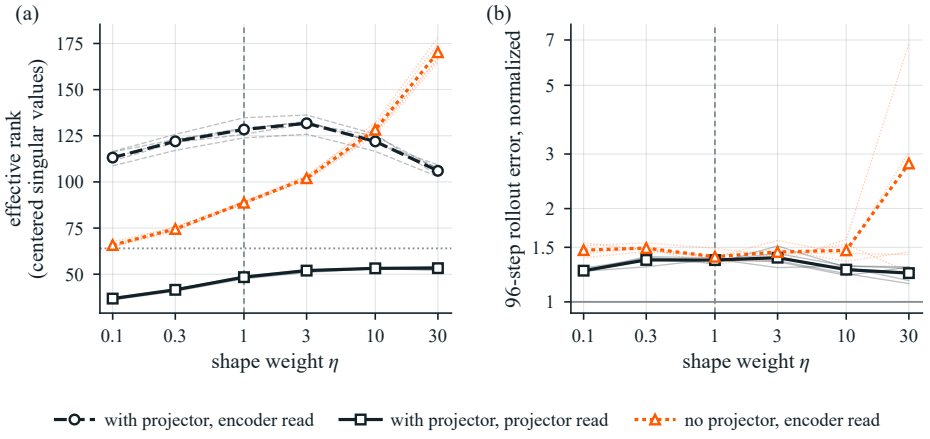

Figure 9: Panels, seeds, schedule and strength as in Figure 6; panel (b) is on a logarithmic axis. The committed endpoint on this dataset failed its own degeneracy gate: every arm, including the unregularized anchor, is worse than a mean predictor, which is why the reference line at 1.0 is drawn (Zhou et al., 2021b).

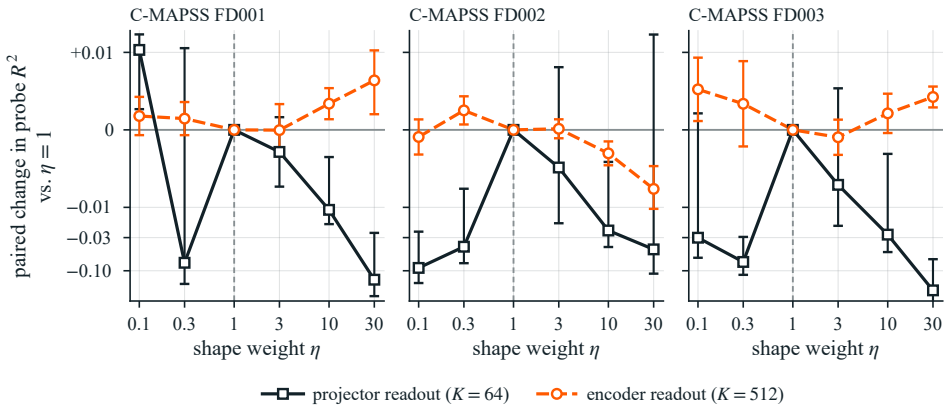

Figure 10: Paired within-seed change in the remaining-life ridge probe R^2 against $\eta = 1$, read at the projector (the regularized layer, width 64) and at the encoder (the deployed backbone, width 512) of the same trained models, five seeds, 95% paired $t(4)$ intervals, vertical axis symmetric-logarithmic below $|0.01|$. On the first C-MAPSS subset the two readouts prefer opposite ends of the grid. On the second and third neither readout beats the default by twice its own measured noise floor, so there is no opposition to reproduce. The forecasting dataset carries no remaining-life readout at all and cannot test the question in either direction. Reported as a non-replication.