# RELIABLE BUT COMPRESSED: LONGER TIMESCALES LOWER DIMENSIONALITY WITH READOUT-DEPENDENT TASK COSTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When a neural representation becomes more reliable and lower-dimensional, does it necessarily become less useful downstream? We test this question in a fixed auditory spiking network by increasing membrane and synaptic timescales while matching mean firing rates. Longer timescales increase reliability to input perturbations but reduce effective dimensionality, producing a reliable-but-compressed representation. Crucially, the apparent downstream cost of this compression is not stable across readout procedures: on the same frozen features, an expanded regularized readout substantially attenuates the long-timescale penalty while improving absolute predictive performance. The geometric change therefore persists, whereas its inferred task cost changes with the evaluation procedure. Mechanistic controls show that compression is much weaker in a separately activity-matched continuous implementation, while slower local state recovery closely follows passive membrane decay. Within this controlled system, these results separate three effects that are easily conflated—representation geometry, perturbation response, and fitted task accessibility—and show that lower dimensionality alone is insufficient to infer an intrinsic loss of downstream utility.

## 1 INTRODUCTION

Changes in neural population geometry are often interpreted together with changes in downstream prediction. If an intervention produces more reliable but lower-dimensional activity and a decoder performs worse, the decline may look like a functional cost of compression. Establishing that interpretation requires separating two sources of change: the representation itself and the finite-data procedure used to fit, select, and calibrate its readout.

Observer dependence in representation evaluation is well established. Predictive $\mathcal{V}$-information, MDL probing, and risk decompositions all distinguish properties of a representation from the function family and training procedure used to evaluate it (Xu et al., 2020; Voita & Titov, 2020; Dubois et al., 2023). We examine a concrete empirical inference that these frameworks leave to be tested: when a controlled dynamical intervention changes representation geometry, does the accompanying task contrast remain supported when exactly the same frozen features are evaluated by another specified readout procedure?

Our assay is a fixed auditory spiking network with parallel fast and slow excitatory–inhibitory reservoirs. We jointly scale membrane and synaptic constants while calibrating population thresholds on training clips to match endpoint mean firing rates. Sounds, connectivity, non-intervened parameters, and model seed remain paired. The intervention produces a *reliable-and-compressed* displacement: input-perturbation reliability $R$ increases while normalized effective dimensionality $D$ decreases. We then freeze the endpoint representations and compare two fully specified linear readout procedures. The expanded regularized procedure improves absolute prediction and changes the long-minus-short accessibility contrast by $+0.05104$ (95% CI $[0.03115, 0.07783]$) relative to the baseline procedure. The functional interpretation therefore changes while the representation does not.

Three results delimit this finding. First, reliable compression survives population-mean activity matching, while its magnitude differs across the tested spiking and continuous implementations. Second, the direct same-feature readout interaction provides no basis for treating the baseline penalty



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.

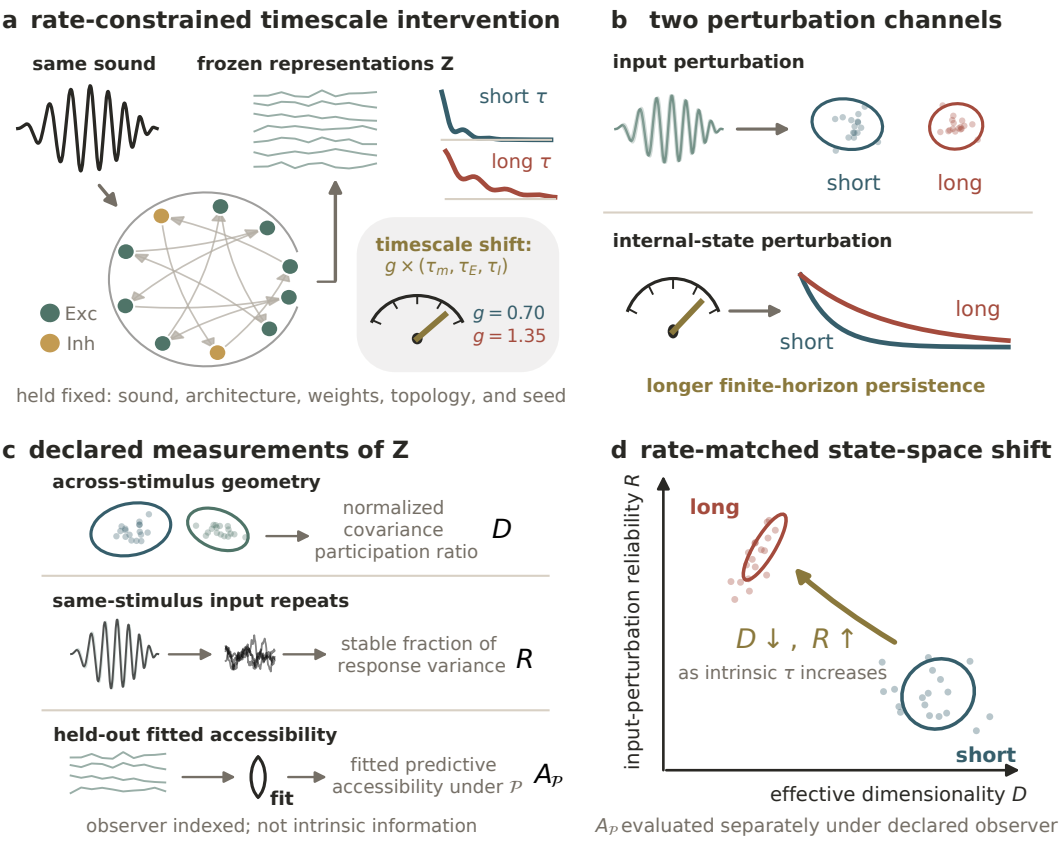



Figure 1: Rate-constrained intervention, two perturbation channels, declared measurements, and operating-point displacement. (a) The same sounds are processed by one fixed recurrent circuit. The intervention factor $g$ jointly scales membrane and synaptic timescales, while population thresholds are calibrated on training clips to match endpoint firing rates; architecture, weights, topology, sounds, and model seed remain paired. (b) Input perturbations measure trial-to-trial response reliability. Matched internal-state kicks measure finite-horizon recovery under the same subsequent input. Longer integration both suppresses the residual effect of input noise and prolongs an internal deviation. (c) The frozen representation $Z$ is evaluated using three explicitly distinct quantities: normalized effective dimensionality $D$; input-perturbation reliability $R = \operatorname{tr} S/(\operatorname{tr} S + \operatorname{tr} N)$, the stable fraction of response variance; and $A_{\mathcal{P}}$, normalized held-out log-loss improvement under declared observer procedure $\mathcal{P}$. (d) After training-only activity matching, longer timescales produce a measured D/R displacement toward lower $D$ and higher $R$. $A_{\mathcal{P}}$ remains a separate observer-indexed evaluation.

as an observer-independent loss of task information. Third, increased reliability to the declared input perturbations coexists with slower local state recovery, whose ordering is numerically accounted for by passive membrane leakage in the tested regime. Together, these results separate geometry, perturbation channel, and fitted task performance within one controlled experiment. They do not identify a unique compression mechanism or a universal tradeoff, and the final readout comparison reuses the test split examined by earlier audits. Figure 1 summarizes the experimental logic.

## 2 Related Work

Representation evaluation is itself observer dependent. Predictive $\mathcal{V}$-information formalizes information relative to a restricted function family (Xu et al., 2020), MDL probing includes the effort required to attain predictive quality (Voita & Titov, 2020), and risk decompositions separate representation usability from probe and encoder generalization (Dubois et al., 2023). These ideas motivate our narrower empirical test: whether a paired timescale effect on task performance changes when the frozen representation is evaluated by another fully specified observer. The empirical increment is the

coupling of a controlled geometric intervention with a direct same-feature readout comparison, rather than observer dependence in isolation.

Liquid-state and reservoir models use recurrent dynamics as temporal feature maps (Maass et al., 2002; Ganguli et al., 2008). Neural timescale hierarchies have been linked to recurrent connectivity and temporal processing (Murray et al., 2014; Chaudhuri et al., 2015; Runyan et al., 2017), while spike reliability depends jointly on stimulus structure and cellular dynamics (Mainen & Sejnowski, 1995; Galán et al., 2008). Work on population dimensionality and representational similarity provides complementary geometric summaries (Gao et al., 2017; Recanatesi et al., 2019; Kornblith et al., 2019). Spiking learning methods provide local or surrogate alternatives to conventional backpropagation (Bellec et al., 2020; Neftci et al., 2019; Zenke & Ganguli, 2018; Kaiser et al., 2020); here the confirmatory backbone is frozen so that learning dynamics do not confound the intervention. These literatures motivate controlling timescales, population activity, geometry, and downstream evaluation in one paired experiment.

# 3 Experimental design

## 3.1 Data and fixed auditory spiking network

We derive an 11-class source/timbre task from FSD50K (Fonseca et al., 2022). The 2,569 two-second clips cover six instruments and five environmental or animal sources. Train, validation, and test contain 1,841, 314, and 414 clips. All clips from a source recording remain in one split, giving zero source overlap.

At 48 kHz, a 64-channel ERB-spaced auditory front end produces parallel envelope and fine-structure features. A local spike encoder drives independent 96-neuron fast and 96-neuron slow LIF reservoirs, each with excitatory and inhibitory units, delayed inhibition, sparse topology-biased recurrence, and separate membrane and synaptic constants. Connectivity and all non-intervened parameters are paired across timescale conditions. The model serves as a controlled dynamical assay with no fitted animal mapping or benchmark claim.

## 3.2 Intervention and evaluation contracts

The focal endpoints multiply membrane and excitatory/inhibitory synaptic time constants by $g \in \{0.70, 1.00, 1.35\}$ at fixed fast–slow separation. The intervention changes more than physical time because refractory steps, input frequencies, clip duration, and effective delays remain on their original scales. For each of 12 fresh network seeds, training clips calibrate population thresholds at short and long endpoints to match the central endpoint's fast and slow mean rates. Thresholds are then frozen. For branch $b \in \{\text{fast}, \text{slow}\}$ and intrinsic constant $k \in \{m, E, I\}$, the intervention and calibration target are

$$\tau_{b,k}(g) = g\tau_{b,k}^{(0)}, \qquad \bar{r}_b^{\text{train}}(g, \theta_b(g)) \simeq \bar{r}_b^{\text{train}}(1, \theta_b(1)) \,. \tag{1}$$

The registered gate requires both populations to remain within 5% on the calibration and held-out sets. This joint timescale–threshold intervention constrains population means only; other rate statistics and synaptic-filter area may change. Every effect is therefore a finite contrast along the calibrated constraint. Appendix A.4 gives the formal constrained derivative and its piecewise-smooth interpretation.

The three coordinates answer different questions: $D$ describes geometric spread, $R$ describes robustness to the declared input perturbation, and $A_{\mathcal{P}}$ describes task access under a declared observer. For centered representation covariance eigenvalues $\lambda_i$, normalized effective dimensionality is

$$D = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} / \min(p, n-1). \tag{2}$$

Input-perturbation reliability is the between-stimulus variance fraction

$$R = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}} + \epsilon}. \tag{3}$$

Writing stable stimulus and residual covariance as $S$ and $N$ gives $\operatorname{logit} R_{\text{tr}} = \log \operatorname{tr} S - \log \operatorname{tr} N$. This decomposition reveals whether stable stimulus power or residual perturbation power drives the

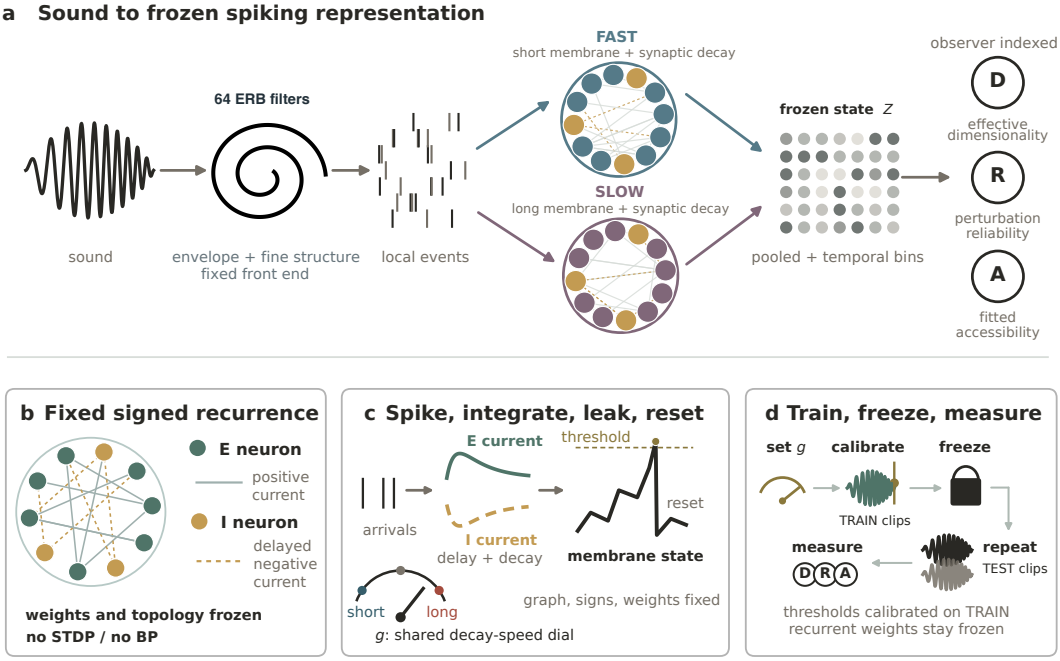



Figure 2: Network architecture and experimental workflow. A 48-kHz waveform is mapped through 64 ERB-spaced filters and parallel envelope/fine-structure features to 128 event channels. Seeded affine maps drive independent 96-neuron fast and slow E/I LIF reservoirs, each with a 75/25 excitatory/inhibitory presynaptic split, sparse frequency-biased recurrence, and delayed inhibition. A global factor $g$ scales membrane and synaptic constants at fixed fast–slow separation. Training clips calibrate fast and slow population thresholds; parameters are then frozen and pooled/eight-bin representations are evaluated for geometry $D$, input-perturbation reliability $R$, internal-state recovery, and observer-indexed fitted accessibility $A_{m,\mathcal{P}}$. All 12 paired seeds share clips, architecture, connectivity, and non-intervened parameters across endpoints.

scalar ratio. To ask how readily task labels can be read from frozen representation $m$, we define probability-sensitive accessibility under procedure $\mathcal{P}$ as

$$\widehat{A}^{\mathrm{info}}_{m,\mathcal{P}} = \frac{\widehat{H}(Y) - \widehat{\mathrm{CE}}_{\mathrm{test}}(\widehat{q}_{\mathcal{P}})}{\widehat{H}(Y)}. \tag{4}$$

It can be negative when fitting fails to attain the empirical prior baseline. Its estimand is procedure-specific predictive accessibility; total mutual information remains unmeasured. Endpoint-wise refitting estimates adapted accessibility $A^{\mathrm{adapt}}_{\mathcal{M},\mathcal{P}}(g)$, whereas applying a decoder fitted at $g_s$ to a target endpoint $g_t$ measures fixed-decoder compatibility $C_{g_s \to g_t}$. Fixed-decoder transfer is reported solely as compatibility and carries no information-loss interpretation.

Confirmatory endpoint contrasts are paired long-minus-short effects, with model seed as the inferential unit, seed-bootstrap 95% intervals, and multiplicity-corrected paired tests. Calibration uses training data. For each declared fit, train/validation choices are frozen before a single endpoint test evaluation. Sequential audits reuse the same test split, as detailed in the appendix. Cross-architecture comparisons respect their actual paired or independent seed provenance. Complete estimands, null hypotheses, and sensitivity analyses are in Appendix B.

## 4 Reliable Compression under Activity Matching

### 4.1 Input reliability increases as representation geometry compresses

Threshold calibration generalized without test-set adjustment. The maximum held-out population-rate mismatch was 0.8997%, and all 48 endpoint–population comparisons reduced the spike-

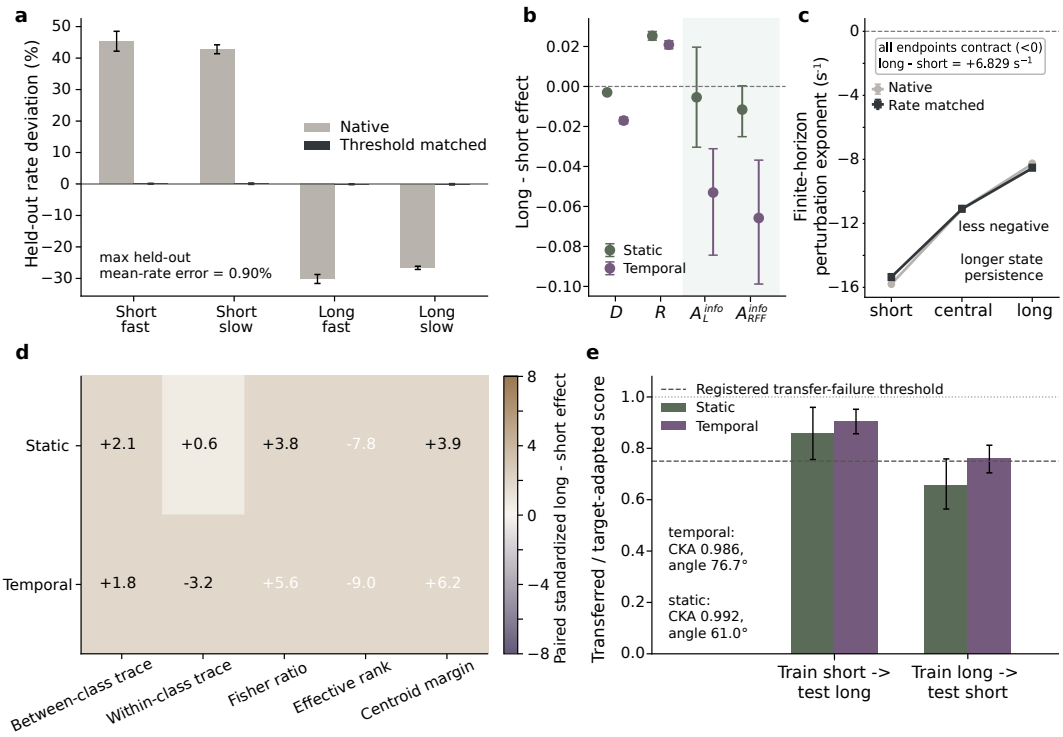



Figure 3: Rate-constrained perturbation-channel dissociation and representational consequences. (a) Training-only threshold calibration removes held-out population-mean differences. (b) Effective dimensionality decreases and input-perturbation reliability increases for static and temporal representations; the shaded accessibility coordinates are baseline-observer diagnostics revisited in Fig. 4. (c) Small membrane-state deviations contract at every endpoint; the finite-horizon perturbation exponent becomes less negative from short to long timescales, indicating longer state persistence. Panel (c) annotates the archived v2.13 endpoint effect ($+6.829\,\mathrm{s}^{-1}$); the forward-only v2.25 audit obtained $+6.912\,\mathrm{s}^{-1}$ under its finite-perturbation protocol and passed the registered reproduction tolerance (Appendix A.3). (d) Effective-rank compression coexists with sharper class-separation statistics. (e) A source-trained fixed probe transfers asymmetrically despite high CKA; retention is its target-endpoint macro-F1 divided by that of a separately target-adapted probe. Bars or intervals in (a,b,e) are 95% model-seed bootstrap intervals, panel (c) shows SEM across 12 paired seeds, and panel (d) shows paired standardized effects. Model seeds are the inferential units.

count Wasserstein distance. After matching, longer timescales reduced $D$ by $0.00297$ (95% CI $[-0.00331, -0.00262]$, Holm $p = 0.00146$) and increased $R$ by $0.02573$ ($[0.02374, 0.02779]$, $p = 0.00146$). The common static macro-F1 effect was $-0.00257$ ($[-0.00869, 0.00498]$, $p = 0.494$), leaving the common accuracy-penalty hypothesis unsupported after mean-rate matching.

The displacement also appears in synchronized eight-bin representations: temporal $D$ changes by $-0.01706$ and temporal $R$ by $+0.02098$. Compression corresponds to approximately $0.57$ fewer participation-ratio dimensions in the 192-dimensional pooled representation and $7.05$ fewer in the eight-bin representation. Compression coexists with sharper class separation: temporal normalized effective rank decreases by $0.08406$, while Fisher trace ratio and normalized centroid margin increase by $0.04937$ and $0.06444$. A finite-repeat correction preserves the ratio-level reliability shift after removing residual variance retained in ten-repeat means (Appendix B). Higher reliability here denotes a lower residual-to-stimulus ratio under the declared perturbation channel.

The following covariance-energy decomposition resolves the geometric change. In the fixed neuron-by-time-bin coordinates, define the diagonal and off-diagonal covariance energies and the variance-

only dimensionality as

$$E_{\mathrm{diag}} = \sum_i C_{ii}^2, \qquad E_{\mathrm{off}} = \sum_{i \neq j} C_{ij}^2, \qquad d_{\mathrm{var}} = \frac{(\sum_i C_{ii})^2}{E_{\mathrm{diag}}}, \qquad \chi = \frac{E_{\mathrm{off}}}{E_{\mathrm{diag}}}. \tag{5}$$

The participation ratio and its paired endpoint displacement then decompose exactly as

$$d_{\mathrm{PR}} = \frac{(\sum_i C_{ii})^2}{E_{\mathrm{diag}} + E_{\mathrm{off}}} = \frac{d_{\mathrm{var}}}{1+\chi}, \qquad \Delta \log d_{\mathrm{PR}} = \Delta \log d_{\mathrm{var}} - \Delta \log(1+\chi). \tag{6}$$

For the long-minus-short contrast, $\Delta \log d_{\mathrm{var}} = -0.105$ and $-\Delta \log(1+\chi) = -0.271$, giving $\Delta \log d_{\mathrm{PR}} = -0.376$. The off-diagonal covariance-energy term therefore contributes 72.0% (95% CI [67.9%, 76.0%]) of the absolute mean displacement (Fig. 5c). This coordinate-dependent algebraic description carries no causal mediation interpretation.

## 5 The inferred task cost changes on the same frozen features

We first fit linear and fixed random-Fourier-feature (RFF) decoders to the same 1,536-dimensional eight-bin temporal representation, with all selection and calibration choices made on validation data. The baseline procedure gives negative long-minus-short accessibility effects for linear ($-0.05284$, 95% CI $[-0.08378, -0.03088]$, Holm $p = 0.00293$) and fixed RFF ($-0.06551$, $[-0.09874, -0.03669]$, $p = 0.00488$), while macro-F1 changes remain near zero. The likelihood and macro-F1 rankings therefore diverge, and the long-endpoint RFF estimate falls below the empirical-prior NLL baseline.

We then keep the raw features and their training-based standardization fixed and use an expanded regularized convex linear procedure. All 12 seeds select the same interior $\ell_2 = 0.1$, and two starts agree within every predictive-function equivalence bound. The long-minus-short accessibility effect becomes $-0.00180$ (95% CI $[-0.00779, 0.00362]$, $p = 0.585$). Comparing the two endpoint contrasts directly, the expanded-minus-baseline change is $+0.05104$ (95% CI $[+0.03115, +0.07783]$, exact $p = 0.000488$). This paired interaction establishes procedure sensitivity without relying on a comparison between one significant and one nonsignificant result.

Absolute prediction improves at both endpoints. Normalized accessibility rises from 0.125 and 0.072 under the baseline procedure to 0.273 and 0.271 under the expanded procedure at the short and long endpoints; macro-F1 rises from 0.390 and 0.383 to 0.452 and 0.459. The attenuation therefore does not arise by making both endpoint predictions less informative. No representation is retrained, so the measured $D/R$ displacement remains unchanged.

This joint contract intervention changes fitting, model selection, and calibration together and does not identify one component as the cause. A PCA sensitivity test likewise changes the endpoint contrast while mapping exactly back to a linear function of the original coordinates (maximum logit error $2.19 \times 10^{-14}$); it diagnoses observer sensitivity without isolating a pure dimensionality mechanism. Additional reproduction, preprocessing, and numerical-solution audits are reported in Appendix D. The baseline effect reproduces under its declared procedure, while the direct interaction shows that it cannot by itself establish decoder-independent loss of label information.

## 6 Mechanistic and implementation boundaries

### 6.1 Input reliability does not imply faster state recovery

A stable scalar filter makes the distinction between perturbation channels explicit:

$$x_{t+1} = a x_t + (1-a)(s + \epsilon_t), \qquad 0 < a < 1, \tag{7}$$

where $s$ has across-stimulus variance $S > 0$ and $\epsilon_t$ is independent input noise with variance $q > 0$. An initial-state deviation obeys $\delta x_k = a^k \delta x_0$, while stationary conditional noise and reliability are

$$N(a) = q\frac{1-a}{1+a}, \qquad R(a) = \frac{S}{S + N(a)}. \tag{8}$$

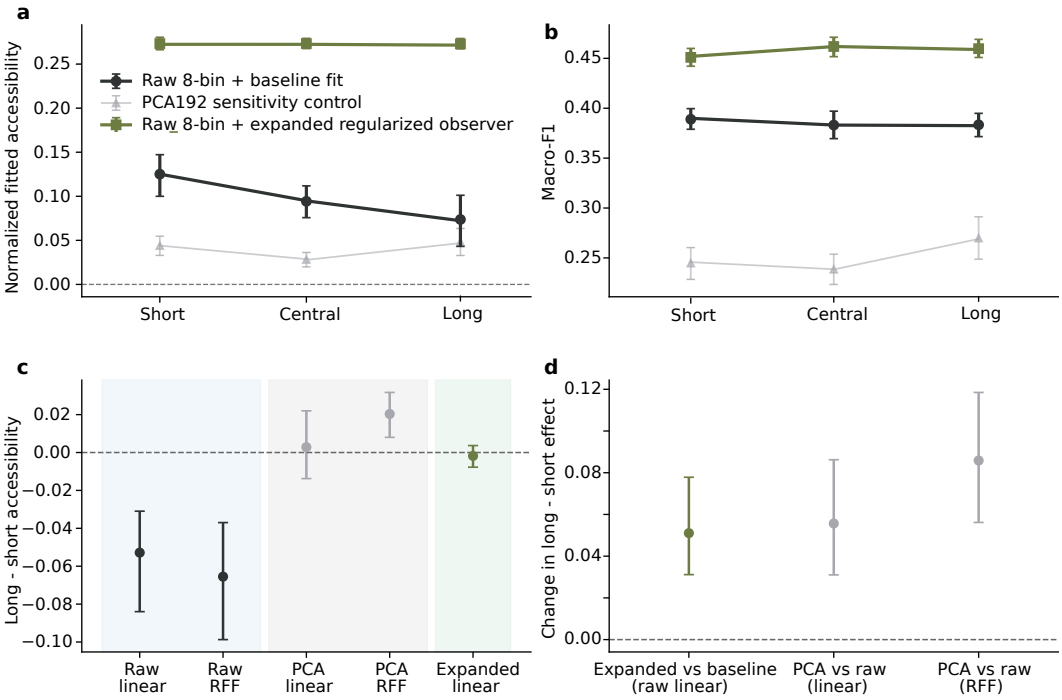



Figure 4: The inferred task contrast depends on the evaluation contract while the representation remains fixed. (a,b) Absolute accessibility and macro-F1 for baseline raw, baseline PCA192, and expanded regularized convex raw linear procedures. PCA192 reduces absolute linear macro-F1 and serves only as a sensitivity control. (c) Long-minus-short accessibility effects under each declared contract. (d) Direct paired interactions quantify how projection and expanded regularized fitting change the endpoint contrast. Error bars are 95% model-seed bootstrap intervals; the 12 paired model seeds are inferential replicates.

Increasing $a$ therefore improves reliability to this input-noise channel while slowing the decay of an internal-state deviation. This example establishes possibility under explicit assumptions and makes no magnitude prediction for the nonlinear recurrent network.

We next perturb the network's internal state. For branch $b$, the measured state is $s^b = (u^b, e^b, i^b)$: membrane potential and excitatory and inhibitory synaptic traces. A membrane perturbation of threshold-normalized amplitude $\epsilon = 10^{-3}$ is applied at clip onset while the other states and future input remain identical. Recovery over $H = 20$ steps is

$$\lambda_H^b = \frac{1}{H\Delta t} \log \frac{\|\delta s^b(H)\|_2}{\epsilon}. \tag{9}$$

We average two fixed directions per clip, take the median across clips within each branch, and conservatively report the less-negative branch. Deviations decay at all endpoints, and longer timescales make the slow-branch exponent less negative by $6.912\,\mathrm{s}^{-1}$ after rate matching. For an event-fixed slow branch with $\tau_m = 100$ ms, passive membrane leakage predicts

$$\Delta\lambda_{\mathrm{leak}} = -\frac{1}{1.35\tau_m} + \frac{1}{0.70\tau_m} = 6.878\,\mathrm{s}^{-1}. \tag{10}$$

The registered effect remains $6.914\,\mathrm{s}^{-1}$ (95% CI $[6.857, 6.962]$) among trajectories whose subsequent spike raster does not change, and it remains $6.944\,\mathrm{s}^{-1}$ ($[6.926, 6.963]$) after recurrence is removed. Only 0.91% of recurrent and 0.55% of recurrence-off slow-branch trajectories change their raster at the registered amplitude. Passive membrane decay therefore accounts numerically for the local ordering without recurrence or threshold-boundary crossings. At larger perturbations, raster changes become common and the architectures diverge. The inference is restricted to local stability and excludes criticality (Appendix A.3). A separate modulation-transfer audit rejects one simple global low-pass prediction and leaves the replacement frequency-domain mechanism unresolved (Appendix F.2).

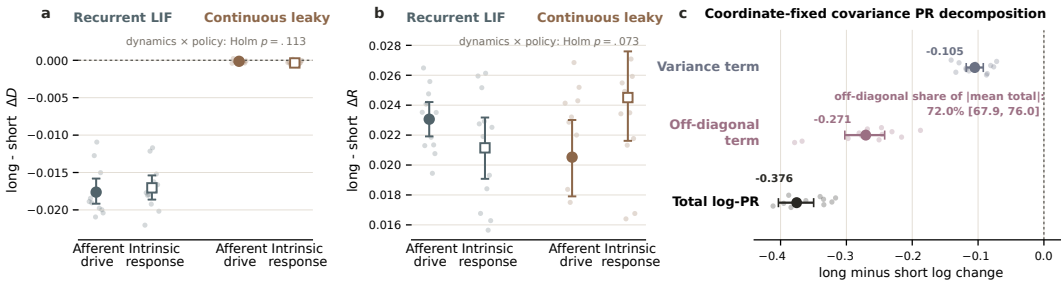



Figure 5: Representational mechanism boundaries. (a,b) Long-minus-short effects for recurrent LIF and continuous leaky dynamics under afferent-drive and intrinsic-response compensation. Points are model seeds; bars are 95% seed-bootstrap intervals. All four cells pass the held-out 5% activity gate, while neither registered dynamics-by-compensation interaction survives Holm correction over $D/R$. (c) Exact log-PR decomposition in the fixed temporal-binned neuron coordinates. The off-diagonal covariance-energy term accounts for 72.0% [67.9%, 76.0%] of the absolute mean displacement. This coordinate-dependent algebraic attribution carries no causal mediation interpretation.

### 6.2 Compression magnitude differs across tested implementations

We repeated the endpoint comparison in two bounded 12-seed controls. The recurrence-off SNN preserves the seeded construction and export-RNG path and passes both rate gates. Its effects ($\Delta D = -0.01731$, $\Delta R = +0.02188$) match the recurrent directions ($-0.01706$, $+0.02100$); paired interactions are $-0.00025$ for $D$ (95% CI $[-0.00153, 0.00110]$) and $+0.00088$ for $R$ ($[-0.00105, 0.00270]$), both Holm $p_H = 1$. These null results leave the effect of recurrence unresolved and provide no equivalence evidence.

The independently initialized, separately activity-calibrated continuous control also increases reliability ($\Delta R = +0.02053$; independent-seed interaction $-0.00047$, 95% CI $[-0.00377, 0.00280]$, Holm $p = 1$). Its $D$ effect is only $-0.000135$ (0.79% of the recurrent magnitude; interaction $+0.01692$, $[0.01526, 0.01850]$, Holm $p = 2.0 \times 10^{-5}$). This tested architecture–compensation package preserves the $R$ direction and strongly attenuates $D$; the design leaves individual component attribution unresolved. Stable and residual traces move oppositely in both SNNs and rise together in the continuous model, so the shared scalar reliability direction arises through different decompositions.

Crossing dynamical implementation with intrinsic-response versus afferent-drive activity compensation retains the $D\downarrow/R\uparrow$ direction in all four cells. Neither registered dynamics-by-compensation interaction survives correction across the $D/R$ family (Holm $p = 0.113$ for $D$ and $0.073$ for $R$; Fig. 5a,b). We find no corrected evidence that the tested compensation policy accounts for the architecture-specific difference in dimensional compression. These controls bound implementation dependence without identifying a unique spiking mechanism. State recovery was not measured in the continuous control.

## 7 Discussion

The main empirical result is that reliable compression and its inferred task cost can be separated within the same controlled system. A paired timescale intervention changes input reliability and representation geometry after mean-activity matching. A subsequent change in readout procedure, with those features held fixed, substantially changes the estimated task contrast and improves absolute predictive quality. The baseline penalty therefore cannot by itself establish a cost intrinsic to the representation. This conclusion rests on a direct paired interaction between procedures.

This experiment provides a concrete instance of the observer dependence emphasized by prior work on representation evaluation. Its additional content is the coupling of a controlled geometric intervention with a same-feature readout comparison: the geometric displacement persists while the functional inference changes. The result motivates reporting both absolute predictive quality and the sensitivity of an intervention contrast to a specified fitting procedure. It does not imply

that compression is generally beneficial, that all task information is preserved, or that the expanded procedure is population optimal.

Two observations bound the mechanistic interpretation. Compression magnitude differs across the tested dynamical implementations, and the comparisons do not isolate a unique spiking mechanism. Improved input reliability also coexists with slower local state recovery, whose ordering closely follows passive membrane leakage in the registered regime. These findings distinguish geometry, input perturbations, and internal recovery without establishing a universal stability or tradeoff law.

The statistical intervals primarily reflect model-seed variation conditional on the supplied recordings and split. Sequential audits reused the same test split, so a frozen comparison on previously unexamined recordings would provide a stronger test of the readout interaction's generality. The intervention also scales synaptic-filter area and integrated recurrent drive, and the continuous control has no biological species mapping. Within these limits, the study establishes an observed procedure dependence in one controlled auditory system: more reliable, lower-dimensional activity does not by itself determine the task cost measured downstream.

### AI USE STATEMENT

OpenAI GPT models and Codex were used to review and debug code, audit analyses, and improve the clarity and language of the manuscript. The authors defined the scientific questions and experimental decisions, executed the experiments, and verified all AI-assisted code, analyses, and manuscript changes against the registered outputs. The authors take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with AI assistance.

### REPRODUCIBILITY STATEMENT

Appendices A–I specify data splits, architecture, interventions, metrics, inference units, reproduction checks, and numerical controls. The anonymous artifact contains frozen protocols, source-hash contracts, per-seed completion markers, derived tables, and cache-only reconstruction commands. FSD50K audio is not redistributed; its public source and group-disjoint metadata are documented in the artifact.

## REFERENCES

Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 11:3625, 2020. doi: 10.1038/s41467-020-17236-y.

Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88(2):419–431, 2015. doi: 10.1016/j.neuron.2015.09.008.

Yann Dubois, Tatsunori Hashimoto, and Percy Liang. Evaluating self-supervised learning via risk decomposition. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8779–8820. PMLR, 2023. URL `https://proceedings.mlr.press/v202/dubois23a.html`.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Roberto F. Galán, G. Bard Ermentrout, and Nathaniel N. Urban. Optimal time scale for spike-time reliability: theory, simulations, and experiments. *Journal of Neurophysiology*, 99(1):277–283, 2008. doi: 10.1152/jn.00563.2007.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008. doi: 10.1073/pnas.0804451105.

Peiran Gao, Eric Trautmann, Byron M. Yu, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, pp. 214262, 2017. doi: 10.1101/214262.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (DECOLLE). *Frontiers in Neuroscience*, 14:424, 2020. doi: 10.3389/fnins.2020.00424.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14 (11):2531–2560, 2002. doi: 10.1162/089976602760407955.

Zachary F. Mainen and Terrence J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995. doi: 10.1126/science.7770778.

John D. Murray, Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17:1661–1663, 2014. doi: 10.1038/nn.3862.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Stefano Recanatesi, Gabriel Koch Ocker, Michael A. Buice, and Eric Shea-Brown. Dimensionality in recurrent spiking networks: Global trends in activity and local origins in connectivity. *PLOS Computational Biology*, 15(7):e1006446, 2019. doi: 10.1371/journal.pcbi.1006446.

Caroline A. Runyan, Eugenio Piasini, Stefano Panzeri, and Christopher D. Harvey. Distinct timescales of population coding across cortex. *Nature*, 548:92–96, 2017. doi: 10.1038/nature23020.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 183–196. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Friedemann Zenke and Surya Ganguli. SuperSpike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, 30(6):1514–1541, 2018. doi: 10.1162/neco_a_01086.

# A WHAT WAS INTERVENED, MATCHED, AND HELD FIXED?

## A.1 ARE TRAIN AND TEST SOURCES INDEPENDENT?

The FSD50K-derived task contains acoustic guitar, electric guitar, organ, piano, trumpet, snare drum, bark, meow, cricket, doorbell, and train. Clips were cropped or padded to two seconds and processed at 48 kHz. There are 714 source-recording groups. Group-disjoint assignment yields 1,841 training, 314 validation, and 414 test clips, with zero source overlap. The nominal uploader group is exactly the source-recording identifier and is not counted as a second leakage control.

## A.2 WHAT STATE DYNAMICS GENERATE THE REPRESENTATION?

The auditory approximation uses 64 ERB-spaced filters over the configured hearing range. Envelope and fine-structure channels are computed in parallel, normalized per clip offline, and converted to binary input events. The fast and slow reservoirs each contain 96 LIF units with a fixed excitatory/inhibitory partition. Recurrent edges obey sparsity, sign, frequency-distance bias, delay, and spectral-normalization constraints, but individual edges are randomly sampled. Fast and slow populations are independent until their features are concatenated at readout.

For branch $b \in \{f, s\}$, the implemented state update is

$$d_t^b = s_{\text{in}} \left( W_{\text{in}}^b x_t + b_{\text{in}}^b \right), \tag{11}$$

$$e_{t+1}^b = \beta_E^b e_t^b + I_t^{b,E}, \qquad i_{t+1}^b = \beta_I^b i_t^b + I_t^{b,I}, \tag{12}$$

$$\widetilde{u}_{t+1}^b = \alpha^b u_t^b + (1 - \alpha^b) \left( d_t^b + e_{t+1}^b - i_{t+1}^b \right), \qquad z_{t+1}^b = \mathbf{1}[\widetilde{u}_{t+1}^b \geq \theta_b], \tag{13}$$

where $\alpha^b = \exp(-\Delta t/\tau_m^b)$ and $\beta_{E/I}^b = \exp(-\Delta t/\tau_{E/I}^b)$. Here $x_t \in \{0,1\}^{128}$ is the encoded envelope/fine-structure event vector, $W_{\text{in}}^b \in \mathbb{R}^{96\times 128}$ is a branch-specific seeded affine projection, $b_{\text{in}}^b$ is initialized to zero, and $s_{\text{in}} = 2$. The input maps are sampled once per model seed, shared across that seed's timescale endpoints, and frozen in the confirmatory assay. The nonnegative currents $I_t^{b,E}$ and $I_t^{b,I}$ collect the excitatory and inhibitory components of the fixed signed delayed recurrence. A spike causes a hard reset to zero and one simulation step of refractoriness. Feedback is disabled. Recurrent matrices are 30% nonzero, use a 25% inhibitory presynaptic population and a frequency-distance kernel of width 0.16, and are row-$\ell_1$ normalized to scale 0.30. Frozen inference uses the hard threshold; the surrogate slope is therefore irrelevant to the reported endpoint measurements.

Table 1: Fixed recurrent assay parameters before the global-timescale intervention and threshold rate matching. Delay values report the effective lag induced by the legacy queue convention.

| Parameter | Fast branch | Slow branch |
|---|---|---|
| Units | 96 | 96 |
| $\tau_m$ (ms) | 20 | 100 |
| $\tau_E$ (ms) | 8 | 24 |
| $\tau_I$ (ms) | 16 | 48 |
| Native threshold | 0.25 | 0.12 |
| Effective inhibitory lag (ms) | 20 | 30 |
| Input scale | 2.0 | 2.0 |
| Reset / refractory | 0 / 10 ms | 0 / 10 ms |

The front end has 128 outputs (64 envelope and 64 fine-structure channels). The local rate encoder is capped at 30 Hz. Clean endpoint features use the configured deterministic rate-matched encoding, whereas noisy repeated-input assays resample events. Offline per-clip normalization precludes claims about online latency or streaming information rate.

The global factor $g$ multiplies membrane and synaptic timescales in both branches while fast–slow separation remains fixed. Every paired seed shares connectivity, input samples, perturbations, and non-intervened parameters. The event increments in the state updates above are not divided by the corresponding kernel area. Thus, for $e_{t+1} = \beta_E e_t + I_t^E$, the discrete impulse area is $(1 - \beta_E)^{-1}$: the registered intervention changes temporal persistence and integrated recurrent drive together. Mean-rate matching does not equate mean synaptic current, synaptic charge, state occupancy, or energy use.

The main simulation step is 10 ms. A corrected 5-ms implementation preserves the physical voltage jump of each encoded event, total inhibitory delays, and refractory duration; Section F.3 reports its endpoint audit.

### A.3 Which perturbation channel does each assay probe?

Input-perturbation reliability is measured on the test split with ten independent responses per clip. For repeat $r$, waveform samples are perturbed as $\widetilde{x}_r = \mathrm{clip}(x + \epsilon_r, -1, 1)$ with $\epsilon_r \sim \mathcal{N}(0, 0.02^2)$ independently across samples, followed by stochastic event encoding. The clean response uses deterministic encoding. Consequently, $R$ measures the combined sensitivity to the declared waveform noise and encoder resampling. Its scope is restricted to this perturbation channel.

The state-recovery audit uses 64 fixed test clips and deterministic encoding. At the start of each clip, two random membrane-potential directions are normalized per sample and scaled to $\epsilon = 10^{-3}$ times the branch threshold. Synaptic currents, delay queues, and refractory counters start from the same values in the reference and perturbed trajectories. The normal threshold, reset, refractory, and recurrent updates are then applied, so later spike events are not clamped to the reference raster. For branch $b$, sample $i$, direction $r$, and horizon $H = 20$, the trajectory statistic is

$$\ell_{ibr}(H) = \frac{1}{H\Delta t} \log \frac{\|\delta s_{ibr}(H)\|_2}{\epsilon}, \tag{14}$$

where $s$ concatenates membrane, excitatory-synaptic, and inhibitory-synaptic states. Directions are averaged before taking the median over samples; the reported pooled value is the less-negative of the fast and slow branch medians. This small-perturbation, finite-horizon recovery statistic has no infinitesimal or global Lyapunov interpretation. A numerical floor is applied to the terminal norm before the logarithm.

For an interval with fixed spike events, the membrane difference obeys $\delta u_{t+1} = \exp[-\Delta t/(g\tau_m)]\delta u_t$. The slow-branch endpoint contrast therefore has the passive baseline $-1/(1.35\tau_m) + 1/(0.70\tau_m) = 6.878\,\mathrm{s}^{-1}$ for $\tau_m = 100$ ms. The registered slow-branch recurrent effect is $6.912\,\mathrm{s}^{-1}$ (95% CI $[6.854, 6.960]$).

We audited amplitudes $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ times threshold and recorded whether any subsequent spike differed from the unperturbed raster. At the registered amplitude $10^{-3}$, the endpoint-averaged slow-branch raster-change fraction is 0.91% with recurrence and 0.55% without it. The recurrent endpoint effect restricted to unchanged-raster trajectories is $6.914\,\mathrm{s}^{-1}$ (95% CI $[6.857, 6.962]$); the recurrence-off effect over all trajectories is $6.944\,\mathrm{s}^{-1}$ ($[6.926, 6.963]$). The changed-raster subgroup contains only six paired seeds with recurrence and two without it, so no subgroup effect is inferred. Across increasing amplitudes, recurrent raster-change fractions are 0.16%, 0.91%, 6.48%, and 49.22%; recurrence-off fractions are 0.03%, 0.55%, 5.50%, and 48.63%. At $10^{-1}$, the all-trajectory endpoint effects diverge ($10.67$ versus $7.32\,\mathrm{s}^{-1}$), marking a sensitivity regime in which event boundaries and architecture matter. A separate rank-48 reduced transition operator is fit with input rank 32 and ridge $10^{-3}$ for the auxiliary operator and resolvent diagnostics; it does not define the trajectory statistic above.

### A.4 What contrast does rate matching identify?

For each seed, the central profile defines target fast and slow population means. Short and long thresholds are calibrated on fixed training clips and then frozen before validation or test evaluation. No test labels, test clips, or synthetic spike correction enter calibration. Both train and held-out population rates must lie within 5% of target. Held-out spike-count Wasserstein distance must improve for every endpoint–population pair. The reported result passes all gates. Threshold adjustment remains an operational model-internal co-intervention with no biological homeostatic fit.

The estimand is the endpoint contrast along the declared calibration policy. It does not identify a threshold-fixed causal effect of an isolated time constant. If $\bar{r}(g, \theta(g)) = \bar{r}_0$, implicit differentiation gives

$$\frac{d\theta}{d\log g} = -\left(\frac{\partial \bar{r}}{\partial \theta}\right)^{-1} \frac{\partial \bar{r}}{\partial \log g}, \tag{15}$$

whenever the local inverse exists. Consequently, the constrained derivative of a smooth reported response $Y$ is

$$\left.\frac{dY}{d\log g}\right|_{\bar{r}} = \frac{\partial Y}{\partial \log g} + \frac{\partial Y}{\partial \theta}\frac{d\theta}{d\log g}. \tag{16}$$

It includes both direct timescale and calibration-policy terms. The intervention also leaves clip duration, stimulus frequencies, refractory steps, and effective delays unscaled, so "longer timescales" should not be read as an exact change of temporal units. For the finite deterministic stimulus set and hard threshold, spike counts can be locally constant and can jump at event boundaries. We therefore use this derivative only as a formal description of a smooth expected-response or surrogate limit. All confirmatory evidence is based on finite endpoint contrasts. We do not claim to have differentiated the hybrid reset dynamics.

# B How are compression, input reliability, and readout measured?

We use $D$, $R$, and $A_{\mathcal{P}}$ as compact names for three different estimands: geometric dimensionality, reliability under a declared input perturbation, and fitted task accessibility under observer $\mathcal{P}$. $A_{\mathcal{P}}$ is neither classification accuracy nor total task information. All reported coordinates are indexed by a measurement contract $\mathcal{M}$. For $D$, it fixes the feature view, centering, covariance estimator, and rank normalization. For $R$, it additionally fixes the perturbation distribution and repeat aggregation. For $A$, the observer $\mathcal{P}$ specifies feature preprocessing, hypothesis class, fitting, selection, and calibration. The notation $\mathbf{o}_{\mathcal{M},\mathcal{P}}$ prevents the relative robustness of $D$ and $R$ in our audits from being mistaken for mathematical invariance.

## B.1 What do the geometry and reliability metrics measure?

For feature covariance $C$ with eigenvalues $\lambda_i$, participation-ratio dimension is $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$. We divide by $\min(p, n-1)$ to compare representations with a common normalized range. This is an effective-rank statistic, not storage capacity or mutual information. For the naive sample covariance this denominator is its rank ceiling. OAS shrinkage generally restores full rank, so its corrected sensitivity report also gives raw PR and PR divided by $p$; the earlier $\min(p, n-1)$ scaling is retained only as a legacy sign-comparison coordinate. Deterministic split-half PR is computed separately in each hash-defined half and normalized by that half's rank ceiling before averaging.

For repeated perturbed responses, reliability is the between-stimulus variance fraction. Repeated data are also decomposed as $Z = \mu(X) + \varepsilon$, with stable stimulus covariance $S = \mathrm{Cov}(\mu(X))$ and residual covariance $N = \mathbb{E}[\mathrm{Cov}(Z \mid X)]$. The trace form

$$R_{\mathrm{tr}} = \frac{\mathrm{tr}\, S}{\mathrm{tr}\, S + \mathrm{tr}\, N} = \frac{1}{1 + \mathrm{tr}\, N / \mathrm{tr}\, S} \tag{17}$$

shows that increased reliability constrains a ratio; it does not imply that absolute residual power or every covariance mode decreases. An equivalent additive decomposition is

$$\mathrm{logit}\, R_{\mathrm{tr}} = \log \mathrm{tr}\, S - \log \mathrm{tr}\, N. \tag{18}$$

We therefore retain $\mathrm{tr}\, S$, $\mathrm{tr}\, N$, their ratio, and clean-to-repeat mean-square displacement as separate diagnostics. With $m = 10$ repeats, the covariance of the repeat mean contains $N/m$. Accordingly, the finite-repeat sensitivity analysis uses

$$\widehat{\mathrm{tr}\, N} = \frac{1}{n}\sum_i \frac{1}{m-1}\sum_r \|Z_{ir} - \bar{Z}_i\|_2^2, \qquad \widehat{\mathrm{tr}\, S} = \mathrm{tr}\, \widehat{\mathrm{Cov}}_i(\bar{Z}_i) - \widehat{\mathrm{tr}\, N}/m. \tag{19}$$

The original plug-in values are retained for provenance, but mechanistic statements use the corrected traces and report if correction changes a sign.

The auxiliary geometry metrics use fixed definitions. With class counts $n_c$, means $\mu_c$, global mean $\mu$, pooled within-class covariance $C_W$, and between-class covariance $C_B = \sum_c (n_c/n)(\mu_c - \mu)(\mu_c - \mu)^\top$, the Fisher trace ratio is $\mathrm{tr}\, C_B / (\mathrm{tr}\, C_W + \epsilon)$. Normalized centroid margin is the mean nearest-centroid distance divided by the root total trace. Normalized effective rank is $\exp[-\sum_i p_i \log(p_i + \epsilon)] / \min(p, n-1)$ for $p_i = \lambda_i / \sum_j \lambda_j$. Fixed-probe retention is the source-trained frozen probe's macro-F1 on the target endpoint divided by the macro-F1 of a separately fitted target-adapted probe on that same endpoint. It is a directional compatibility measure with no information-preservation interpretation.

### B.2 WHY IS FITTED ACCESSIBILITY OBSERVER DEPENDENT?

For decoder family $\mathcal{F}$, population-optimal accessibility is

$$A^{*,\mathrm{info}}_{m,\mathcal{F}} = \frac{H(Y) - \inf_{q\in\mathcal{F}} \mathrm{CE}(q)}{H(Y)}. \tag{20}$$

If the family contains the true-prior predictor, this quantity is nonnegative. The experiment instead estimates a fitted score under a finite procedure $\mathcal{P}$ specifying scaling, feature map, initialization, optimization, hyperparameter selection, early stopping, and calibration. Its variational gap is

$$I(Y;Z) - [H(Y) - \mathrm{CE}(q)] = \mathbb{E}_Z \,\mathrm{KL}(p(Y \mid Z) \| q(Y \mid Z)) \geq 0. \tag{21}$$

A fitted score can therefore be negative and can change because the approximation or estimation gap changes. Exact mutual information remains unmeasured.

We distinguish two functionally different estimands. In adapted accessibility, the declared procedure is fitted separately at each endpoint and asks how well that endpoint can support a newly adapted observer. In fixed-decoder compatibility, the source-endpoint decision function is frozen and evaluated at another endpoint. The latter mixes coordinate alignment with task performance and measures decoder compatibility only.

### B.3 WHAT IS THE BASELINE OBSERVER?

The reported normalization uses the empirical entropy of the same 414 test labels, $H_{\mathrm{test}}(Y) = 2.278$ nats; the corresponding empirical-prior cross-entropy has $\widehat{A}^{\mathrm{info}}_{m,\mathcal{P}} = 0$ by construction. Each condition is standardized from its training split. The canonical rule replaces training standard deviations below $10^{-6}$ by one. A fixed RFF probe concatenates standardized linear coordinates with

$$\phi_j(z) = \sqrt{2/512}\cos\left((\gamma/\sqrt{p})w_j^\top z + b_j\right), \quad w_j \sim \mathcal{N}(0, I), \quad b_j \sim \mathcal{U}(0, 2\pi), \tag{22}$$

and standardizes those 512 random features from training data. The multiplier $\gamma \in \{0.5, 1, 2\}$ is selected on the central validation split and shared across conditions within a seed, as is the random map. Linear softmax weights are fit by full-batch AdamW for at most 160 epochs (learning rate 0.03, decoupled weight decay $10^{-4}$), retaining the lowest validation cross-entropy with patience 20 and minimum improvement $10^{-5}$. A final temperature is selected on validation data from $\{0.25, 0.5, 0.75, 1, 1.5, 2, 3, 4\}$ before one test evaluation. The representation and all SNN parameters remain frozen.

### B.4 HOW IS THE EXPANDED REGULARIZED READOUT FITTED?

The gauge-fixed raw-coordinate analysis keeps the baseline 1,536-dimensional concatenated eight-bin representation and its condition-specific training standardization fixed. To remove the common-logit gauge, multiclass weights and biases are parameterized as $W = HB$ and $b = Ha$, where $H \in \mathbb{R}^{11\times 10}$ is an orthonormal Helmert contrast basis. For standardized training features $z_i$, the optimized objective is

$$\mathcal{L}_\lambda(B, a) = \frac{1}{n}\sum_{i=1}^{n} -\log \mathrm{softmax}(H(Bz_i + a))_{y_i} + \frac{\lambda}{2}\|HB\|_F^2. \tag{23}$$

This is mean cross-entropy plus an $\ell_2$ penalty on weights in the standardized raw-feature coordinates; biases are not penalized. For each model seed, $\lambda$ is selected only at the central endpoint by minimizing central validation NLL over $10^{-4}, \ldots, 10^2$ plus geometric refinement between the neighboring coarse values. The selected value is then shared across short, central, and long endpoints. After each endpoint fit, temperature is selected on that endpoint's validation split from 145 log-spaced values over $[0.125, 64]$. Each endpoint test split is evaluated once after that fit's train/validation choices are frozen. Sequential audits reuse the same data split, model seeds, and representation caches. The final test set was therefore visible to the sequential research program even though each fit froze its declared choices before endpoint evaluation. Every seed selects the interior value $\lambda = 0.1$; all selected temperatures are interior. Float64 strong-Wolfe L-BFGS is run from both a PCA-backmapped and a zero initialization. The two starts test numerical solution invariance under this one declared convex objective. Endpoint-representation equivalence remains outside this test.

### B.5 What is the independent unit in each comparison?

The independently initialized model seed is the primary unit. Within an architecture, paired bootstrap confidence intervals resample seedwise long-minus-short effects and exact sign-flip tests preserve synchronized metrics within a seed. The recurrence-off SNN follows the same seeded SNN construction and export-RNG path as the recurrent SNN, so their interaction is also paired by model seed. The continuous leaky implementation independently draws weights, sparsity masks, spectral normalization, and perturbations; equal integer seed labels are therefore not common random numbers. Its interaction with the recurrent SNN uses an independent bootstrap and 200,000 two-sided random-label permutations. Holm correction is applied separately to the three matched endpoint metrics, the six view-by-probe endpoint contrasts, the two projection-by-timescale interactions, the two scaler-by-timescale interactions, and the two absolute implementation- discrepancy contrasts. The four architecture $D/R$ interactions form one additional Holm family. With 12 paired seeds, the smallest two-sided exact sign-flip value is $2/2^{12} = 0.000488$; the independent permutation estimate has resolution $1/200001$. Clip and perturbation repeats are never used as independent inferential replicates. The two dynamics-by-compensation interactions form a separate Holm family over $D$ and $R$. These intervals quantify initialization variability for one fixed dataset and split; they do not cover new datasets, alternative split assignments, or architecture families.

## C What is the reliable-and-compressed endpoint shift?

Table 2: Long-minus-short effects for native and rate-matched pooled features.

| Condition | $\Delta D$ | $\Delta R$ | $\Delta$ macro-F1 |
|---|---|---|---|
| Native | $-0.00253$ | $+0.01125$ | $-0.01388$ |
| Rate matched | $-0.00297$ | $+0.02573$ | $-0.00257$ |
| Rate-matched 95% CI | $[-0.00331, -0.00262]$ | $[0.02374, 0.02779]$ | $[-0.00869, 0.00498]$ |
| Rate-matched Holm $p$ | $0.00146$ | $0.00146$ | $0.494$ |

Table 3: Rate-matched covariance-geometry effects.

| Metric | Pooled static | Eight-bin temporal |
|---|---|---|
| Between-class trace | $+0.00458$ | $+0.03340$ |
| Within-class trace | $+0.00195$ (inconclusive) | $-0.14102$ |
| Fisher trace ratio | $+0.05481$ | $+0.04937$ |
| Normalized effective rank | $-0.01606$ | $-0.08406$ |
| Normalized centroid margin | $+0.04585$ | $+0.06444$ |
| Trace reliability | $+0.02542$ | $+0.02098$ |

Low–high CKA remains 0.9917 for pooled static and 0.9861 for temporal features. Source-scaled short-to-long/long-to-short fixed-probe retention is 0.858/0.658 for static and 0.906/0.762 for temporal features. Both directions do not satisfy the registered broad-failure ceiling, so the transfer result supports directional coordinate dependence within the tested endpoints.

## D Does the accessibility cost survive changes in the observer?

### D.1 Does the baseline accessibility effect reproduce?

The audit uses the original v2.14 matched caches for `low_matched`, `central_reference`, and `high_matched` across seeds 21101–21112. It pins hashes for the original fitting source, its configuration, the independent v2.16 source and configuration, and both archived result tables. Resume markers bind those inputs to artifact hashes.

Stage 1 calls the original fitting entrypoint and a common adapter on the identity representation. The adapter tolerance is $10^{-7}$ and the archived-score tolerance is $10^{-4}$, with exact metadata checks. All 12 seeds must pass before Stage 2. Stage 2 fits PCA only on central training features, applies the map

unchanged to all endpoints and splits, and calls the same original probe function. A one-bin mean provides a second 192-coordinate view. Stage 3 reproduces the independent v2.16 baseline before changing its near-zero-variance scaling rule. The SNN remains frozen throughout.

Table 4: v2.17 long-minus-short fitted accessibility effects. Intervals are 95% seed-bootstrap CIs; $p_H$ is Holm-adjusted across six contrasts.

| View | Probe | Effect | 95% CI | $p_H$ |
|---|---|---|---|---|
| Identity 8-bin | Linear | $-0.05284$ | $[-0.08378, -0.03088]$ | .00293 |
| Identity 8-bin | Fixed RFF | $-0.06551$ | $[-0.09874, -0.03669]$ | .00488 |
| PCA192 8-bin | Linear | $+0.00283$ | $[-0.01378, 0.02206]$ | 1.000 |
| PCA192 8-bin | Fixed RFF | $+0.02034$ | $[0.00802, 0.03172]$ | .03906 |
| One-bin 192-D | Linear | $-0.00625$ | $[-0.03389, 0.01692]$ | 1.000 |
| One-bin 192-D | Fixed RFF | $-0.02548$ | $[-0.04416, -0.00596]$ | .10400 |

The identity result establishes reproducibility within the declared procedure. The PCA result establishes lack of invariance under the tested projection and leaves a pure dimensionality mechanism unidentified. The one-bin result jointly changes temporal content and dimension. No row is interpreted as exact mutual information.

The confirmatory comparison is the within-seed interaction

$$\Delta_{\text{int}} = (A_{\text{long}} - A_{\text{short}})_{\text{modified}} - (A_{\text{long}} - A_{\text{short}})_{\text{reference}}. \tag{24}$$

It avoids interpreting a difference between separate significance decisions.

Table 5: Direct paired evaluation-contract interactions. PCA192 is compared with the identity eight-bin view; expanded fitting is compared on the same raw features; and the safe scaler is compared with the legacy $10^{-6}$ floor. Projection and scaler rows retain their registered within-family Holm adjustment; the single expanded-fit follow-up reports its exact unadjusted sign-flip $p$.

| Intervention | Probe | $\Delta_{\text{int}}$ | 95% CI | $p$ |
|---|---|---|---|---|
| PCA192 minus identity | Linear | $+0.05567$ | $[0.03101, 0.08622]$ | .00195 |
| PCA192 minus identity | Fixed RFF | $+0.08585$ | $[0.05616, 0.11849]$ | .00195 |
| Expanded minus original fit | Linear | $+0.05104$ | $[+0.03115, +0.07783]$ | 0.000488 |
| Safe minus legacy scaler | Linear | $-0.12695$ | $[-0.64437, 0.40740]$ | 1.000 |
| Safe minus legacy scaler | Fixed RFF | $-0.07699$ | $[-0.57105, 0.37826]$ | 1.000 |

Table 6: Absolute test metrics for the original identity and PCA192 views and the expanded-convex identity procedure. The empirical-prior NLL baseline is 2.278 nats and has normalized accessibility zero. Values are means over 12 model seeds.

| View | Probe | Endpoint | $\widehat{A}^{\text{info}}_{m,\mathcal{P}}$ | NLL | Macro-F1 |
|---|---|---|---|---|---|
| Identity | Linear | Short | .125 | 1.993 | .390 |
| Identity | Linear | Central | .094 | 2.063 | .383 |
| Identity | Linear | Long | .072 | 2.113 | .383 |
| Identity | Fixed RFF | Short | .051 | 2.162 | .361 |
| Identity | Fixed RFF | Central | .034 | 2.200 | .364 |
| Identity | Fixed RFF | Long | $-.015$ | 2.311 | .363 |
| PCA192 | Linear | Short | .044 | 2.178 | .246 |
| PCA192 | Linear | Central | .028 | 2.213 | .239 |
| PCA192 | Linear | Long | .047 | 2.171 | .269 |
| PCA192 | Fixed RFF | Short | .077 | 2.103 | .279 |
| PCA192 | Fixed RFF | Central | .075 | 2.108 | .285 |
| PCA192 | Fixed RFF | Long | .097 | 2.057 | .302 |
| Expanded convex identity | Linear | Short | 0.273 | 1.655 | 0.452 |
| Expanded convex identity | Linear | Central | 0.273 | 1.655 | 0.462 |
| Expanded convex identity | Linear | Long | 0.271 | 1.659 | 0.459 |

Table 7: Endpoint effects in the v2.18 affine-backmapping and convex-refit audit. Confidence intervals and exact sign-flip tests use model seed as the independent unit.

| Estimator | long−short | 95% CI | $p$ |
|---|---|---|---|
| Exact PCA function, backmapped | $0.00283$ | $[-0.01389, 0.02179]$ | $0.794$ |
| Raw-coordinate convex refit | $-0.00415$ | $[-0.01417, 0.00593]$ | $0.446$ |
| Refit minus exact backmap | $-0.00698$ | $[-0.02954, 0.01340]$ | $0.583$ |

### D.2 Can preprocessing explain cross-implementation score differences?

The canonical scaler replaces training standard deviations below $10^{-6}$ with one. The independent implementation clamps them to $10^{-6}$, which can amplify a coordinate that is constant in training but nonzero on held-out data. A synthetic contract test confirms the possible amplification factor of $10^6$ and serves only as a software-contract check. In the formal data, changing the two legacy scaling sites jointly reduces absolute NLL disagreement by $3.618$ nats for linear probes (95% CI $[2.522, 4.890]$) and $3.378$ for fixed RFF ($[2.469, 4.294]$), both Holm $p = 0.00098$. Scaling therefore makes a substantial contribution to the implementation difference while leaving other sources unresolved.

The ranking-specific test is negative. The safe-minus-legacy change in the long-minus-short effect is $-0.12695$ for linear and $-0.07699$ for fixed RFF; both intervals include zero and both Holm-adjusted $p$ values equal one (Table 5). Near-zero-variance training coordinates are present in every endpoint (means 219, 243, and 265 of 1,536 at short, central, and long); the long-minus-short count change is $+46.33$ with all 12 paired seeds in the same direction. Yet held-out activation counts do not increase monotonically, and their descriptive association with the scaler interaction provides no mediation evidence. The scaler rule is a verified source of absolute cross-implementation error; the endpoint ranking remains unexplained.

### D.3 Does projection create a new linear function, and does refitting preserve the ranking?

For each model seed and timescale profile, we analytically composed the fitted PCA-192 linear probe with the centering and projection operators. The resulting affine classifier acts directly on the original 1,536-dimensional concatenated eight-bin feature vector. Its logits match the PCA-space classifier to numerical precision, with maximum absolute error $2.19 \times 10^{-14}$. This is an exact function-equivalence test: every fitted PCA-space linear classifier can be implemented exactly in the original coordinates. PCA therefore preserves linear expressibility for the fitted function. Raw-coordinate direction preservation and equal finite-sample generalization under refitting remain outside this result.

Raw-coordinate refitting substantially improves absolute accessibility. The initial v2.18 convergence gate fails across all 12 seeds: the selected $\ell_2$ value is the upper boundary of the original grid ($10^{-3}$), and the maximum test-probability difference between zero- and backmap-initialized fits is $0.0251$. The regularized objectives are nearly identical, which excludes nonconvexity as the explanation and identifies an incomplete regularization contract through the boundary optimum.

We prospectively froze a v2.19 cache-only closure before inspecting its results. It uses an orthonormal Helmert contrast basis to remove the softmax common-shift gauge, searches $\ell_2$ from $10^{-4}$ through $10^2$ with local geometric refinement, searches 145 temperatures over $[0.125, 64]$, and uses repeated float64 strong-Wolfe LBFGS blocks. Both PCA-backmapped and zero initializations are fitted independently for every seed and endpoint.

The expanded closure reduces the worst probability discrepancy by a factor of 8,821 relative to v2.18. The backmap-initialized long-minus-short effect is $-0.00180$ (95% CI $[-0.00779, 0.00362]$, exact $p = 0.585$), while the zero-minus-backmap endpoint interaction is $2.54 \times 10^{-9}$ (95% CI $[-1.88 \times 10^{-8}, 2.78 \times 10^{-8}]$). We therefore distinguish a supported functional-equivalence result from the failed strict omnibus certificate. This closes the multiple-optima alternative at the predictive function level without claiming a complete first-order certificate or a biological decoder.

Table 8: Registered v2.19 convex-convergence gates. The strict conjunction is false because the maximum gradient narrowly exceeds its threshold; every predictive-function and endpoint-equivalence gate passes.

| Gate | Observed maximum | Threshold | Result |
|---|---|---|---|
| Interior $\ell_2$ | 0.1 for 12/12 | interior | Pass |
| Interior temperature | 1.477–1.610 | $(0.125, 64)$ | Pass |
| Gradient | $1.394 \times 10^{-7}$ | $10^{-7}$ | Fail |
| Objective difference | $2.465 \times 10^{-12}$ | $10^{-7}$ | Pass |
| Centered-logit difference | $3.855 \times 10^{-5}$ | $10^{-3}$ | Pass |
| Probability difference | $2.841 \times 10^{-6}$ | $10^{-4}$ | Pass |
| Gauge-fixed parameter difference | $1.793 \times 10^{-5}$ | $10^{-3}$ | Pass |
| Endpoint-effect difference | $1.132 \times 10^{-7}$ | $10^{-3}$ | Pass |

Table 9: Descriptive long-minus-short endpoint effects under the v2.18 scaler-site factorial. "Either" indicates that both post-RFF rules give the same result.

| Readout | Scaling condition | Effect | $p$ |
|---|---|---|---|
| Linear | Legacy input | 0.09352 | 0.747 |
| Linear | Unit-fallback input | −0.03344 | 0.0288 |
| Fixed RFF | Legacy input, either post rule | 0.02211 | 0.938 |
| Fixed RFF | Unit input, either post rule | −0.05488 | 0.0400 |
| Linear | Unit input, risk coordinates deleted | −0.04477 | 0.00098 |

### D.4 Which scaling site changes absolute scores?

The scaler audit separately varied the input scaler and the post-RFF scaler. The post-RFF intervention has exactly zero effect under both input-scaling rules. Input scaling strongly changes absolute scores, yet its interaction with the long–short endpoint contrast is unsupported for linear and fixed-RFF readouts. Table 9 reports the descriptive endpoint effects; the registered interaction tests provide the inferential claim.

The direct input-scaler interactions are $-0.12695$ for the linear probe ($p = 0.655$) and $-0.07699$ for fixed RFF ($p = 0.772$); the post-RFF interaction is exactly zero. Risk-coordinate deletion is descriptive because a direct deletion-by-endpoint interaction was not registered. Collectively, the audit localizes the absolute implementation discrepancy to input scaling. Scaler mediation of the timescale-dependent ranking remains unsupported.

## E Which $D/R$ effects appear under changes in dynamical implementation?

The two controls preserve dataset, splits, endpoint factors, temporal binning, perturbation distribution, and measurement estimators. Cross-model random realizations remain independent. The recurrence-off SNN changes only `use_recurrent`, follows the same seeded construction and export-RNG path as the recurrent SNN, and repeats the training-only threshold calibration; 60/60 endpoint runs pass train and held-out rate gates. The continuous control uses two 96-unit recurrent tanh leaky reservoirs. For branch $b$, its implemented update is

$$c_{b,t} = \tanh\big(\kappa_b W_b^{\text{in}} x_t + W_b^{\text{rec}} h_{b,t}\big), \qquad h_{b,t+1} = \alpha_b h_{b,t} + (1 - \alpha_b) c_{b,t}, \tag{25}$$

where $\alpha_b = \exp[-\Delta t/(g\tau_b)]$. Input matrices have independent $\mathcal{N}(0, 1/d_{\text{in}})$ entries. Recurrent matrices have independent $\mathcal{N}(0, 1/d_b)$ entries, a Bernoulli sparsity mask, and are rescaled to the registered spectral radius; states start at zero. Pooled and eight-bin representations average $h_t$ over time. Training-only input-gain calibration matches branchwise mean absolute activity; all 12 seeds pass the registered 10% gate and the maximum mismatch is 4.235%. This activity match constrains mean absolute state under a compensation policy that differs from threshold calibration; firing-rate and state-distribution equality remain unconstrained. The downloaded summary artifacts support

the pinned construction/RNG provenance. Byte-level identity of the original server caches remains unverified.

Table 10: Long-minus-short temporal endpoint effects across bounded architecture controls. Intervals are 95% seed-bootstrap CIs. Leaky-reservoir accessibility is descriptive because its initialization-invariance gate fails.

| Architecture | $\Delta D$ [95% CI] | $\Delta R$ [95% CI] |
|---|---|---|
| Recurrent SNN | $-0.01706$ $[-0.01865, -0.01539]$ | $+0.02100$ $[0.01898, 0.02302]$ |
| Continuous leaky | $-0.000135$ $[-0.000237, -0.000028]$ | $+0.02053$ $[0.01787, 0.02299]$ |
| No-recurrence SNN | $-0.01731$ $[-0.01918, -0.01560]$ | $+0.02188$ $[0.02117, 0.02261]$ |

Raw-linear accessibility effects are respectively $-0.00180$, $+0.01055$, and $-0.00194$. The continuous value is descriptive because its initialization probability gap $3.06 \times 10^{-4}$ exceeds the registered $10^{-4}$ gate.

Table 11: Provenance-correct cross-architecture interactions. Each entry is the target architecture's long-minus-short effect minus the recurrent SNN effect. The recurrence-off comparison is paired by model seed; the independently initialized continuous comparison uses independent bootstrap intervals and 200,000 random-label permutations. That permutation test uses the stronger exchangeability null of equal effect distributions. A studentized Welch-permutation sensitivity analysis, reported in the generated P0/P1 audit, targets equality of means while allowing unequal variances. Holm $p$ values cover the four primary $D/R$ interactions.

| Target minus recurrent | Metric | Interaction | 95% CI | Holm $p$ |
|---|---|---|---|---|
| Continuous leaky | $D$ | $+0.01692$ | $[0.01526, 0.01850]$ | $2.0 \times 10^{-5}$ |
| Continuous leaky | $R$ | $-0.00047$ | $[-0.00377, 0.00280]$ | 1.000 |
| No recurrence | $D$ | $-0.00025$ | $[-0.00153, 0.00110]$ | 1.000 |
| No recurrence | $R$ | $+0.00088$ | $[-0.00105, 0.00270]$ | 1.000 |

The three dimensionality estimators (naive, OAS-shrinkage, and deterministic split-half participation ratio) agree in sign within every architecture. OAS is also reported with full-rank normalization in the generated sensitivity table. Sign agreement provides no proof of unbiased dimension recovery or magnitude replication: the absolute leaky-reservoir $D$ effect is 0.79% of the recurrent SNN effect, while its $R$ effect is 97.7% as large. The undetected recurrence-off interactions also provide no equivalence evidence. These controls establish that the same direction appears after recurrent drive is disabled and that the separately activity-calibrated continuous architecture–compensation protocol, under its own matching policy, reproduces the reliability direction. The source of attenuated dimensional compression remains unresolved among continuous dynamics, connectivity normalization, compensation policy, and other implementation differences.

### E.1 Does compensation policy explain the architecture difference?

To separate compensation policy from dynamical implementation, we crossed recurrent LIF and continuous leaky dynamics with intrinsic-response and afferent-drive activity compensation. Intrinsic-response compensation changes the LIF threshold or the continuous model's complete-preactivation gain; afferent-drive compensation changes only the branchwise input multiplier. Each endpoint matches its own implementation's central branchwise first moment (spike probability per step for LIF, mean absolute state for the continuous model); activity units are not equated across implementations. Both recurrent-LIF policies were rebuilt under one runtime and shared the same central representation; all four cells passed the held-out 5% activity gate. The recurrent-LIF endpoint effects were $\Delta D = -0.01762$, $\Delta R = +0.02306$ under afferent-drive compensation and $-0.01705$, $+0.02115$ under intrinsic-response compensation. The corresponding continuous-leaky effects were $-0.000135$, $+0.02053$ and $-0.000355$, $+0.02451$. The dynamics-by-compensation interaction was $+0.000794$ for $D$ (95% CI $[-0.000137, 0.001658]$, Holm $p = 0.113$) and $-0.00590$ for $R$ ($[-0.01118, -0.00105]$, unadjusted $p = 0.0366$, Holm $p = 0.0732$). Thus neither interaction survives the registered family correction. The registered tests do not support compensation choice alone as an explanation for the architecture-dependent dimensionality magnitude. The reliability interaction remains suggestive and lacks confirmatory support.

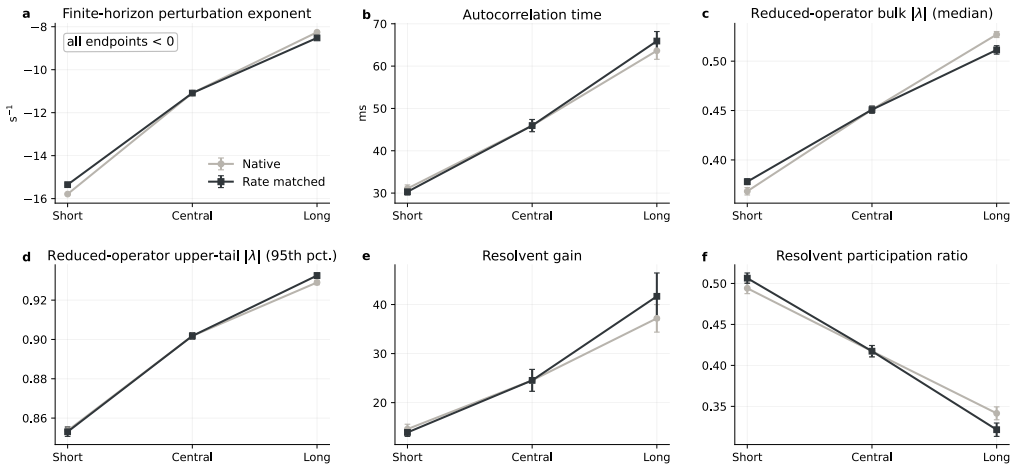



Figure 6: Full finite-horizon dynamical audit across timescale endpoints. Native and rate-matched short–central–long curves share the central endpoint; comparisons are within intervention. Panel (a) shows the small membrane-state perturbation statistic defined above; the remaining panels show autocorrelation, reduced-operator spectra, resolvent gain, and resolvent participation. Error bars are SEM across 12 paired model seeds. These auxiliary diagnostics remain descriptive with respect to exact criticality.

Reliability also has architecture-dependent components. In the recurrent and recurrence-off SNNs, stable trace increases while residual trace decreases. In the continuous leaky reservoir, both increase, with a larger proportional rise in stable power. Coordinate-scale differences preclude raw-trace magnitude comparisons across architectures. This decomposition demonstrates why higher trace reliability cannot be equated with universal noise suppression.

# F Additional dynamical boundary checks

## F.1 Auxiliary linear and spectral accounts

For a stable auxiliary linear system

$$x_{t+1} = Jx_t + Bs_t + C\eta_t, \qquad \rho(J) < 1, \tag{26}$$

an initial-state deviation evolves as $J^k \delta x_0$; stationary input-perturbation covariance is

$$N = \sum_{k=0}^{\infty} J^k C Q C^\top (J^\top)^k. \tag{27}$$

The second quantity depends jointly on input coupling, orientation, stimulus covariance, transient amplification, and contraction rate. This auxiliary model extends the possibility argument beyond one dimension and carries no fitted-linearization or causal-mediation interpretation.

A full-spectrum majorization condition is sufficient for lower participation ratio. The mean-spectrum gate fails despite 11/12 seeds satisfying the ordering. A joint scale-plus-probe-angle model predicts little held-out transfer loss (leave-one-seed-out $R^2 = 0.0558$, permutation $p = 0.538$). Neither low-order account is promoted as the mechanism.

## F.2 Is the shift a simple low-pass effect?

A no-readout modulation-transfer assay uses shared broadband carriers at 2, 4, 8, 12, 20, and 32 Hz. The prediction that longer timescales lower combined modulation bandwidth is contradicted: the observed global-scale coefficient is positive. The prediction that fast–slow separation increases branch bandwidth separation is unsupported. These results reject the two preregistered frequency-domain predictions and leave other low-pass contributions and the alternative MTF mechanism unresolved.

### F.3 Does the endpoint result survive a smaller simulation step?

An initially attempted 5-ms audit changed event-current jump and effective delay conventions and correctly failed its gate. The corrected audit preserves physical event impact and delays. Across the same five seeds, central fast and slow rates differ from the 10-ms reference by 4.32% and 4.51%; $D$ and $R$ effect magnitudes differ by 5.69% and 1.79%, with directions preserved. The direction predicted for the preregistered modulation-bandwidth statistic remains contradicted. Other frequency-selective filtering, nonlinear spike transformations, and readout-induced temporal averaging remain possible. The result establishes endpoint numerical agreement only; full-grid convergence remains untested.

## G Finite-repeat, covariance, and calibration audit

On the clean test-set temporal-binned representation, with $m = 10$ perturbation repeats, we correct the plugin residual trace by $m/(m-1)$ and subtract $N/m$ from the covariance of repeat means. The recurrent-SNN high-minus-low reliability effect remains 0.0233 (95% CI [0.0211, 0.0256]): corrected stable trace changes from 0.5314 to 0.5876, and unbiased residual trace from 2.0228 to 1.9511. The dimensionality sign is unchanged under naive, OAS full-rank-normalized, and split-half estimators.

Let $C$ be the sample covariance in the fixed 1,536-dimensional neuron-by-time-bin basis. Define $d_{\mathrm{var}} = (\sum_i C_{ii})^2 / \sum_i C_{ii}^2$ and $\chi = \sum_{i \neq j} C_{ij}^2 / \sum_i C_{ii}^2$. Because $\mathrm{tr}(C^2) = \sum_i C_{ii}^2 + \sum_{i \neq j} C_{ij}^2$, $d_{\mathrm{PR}} = d_{\mathrm{var}}/(1+\chi)$ exactly. Across paired seeds, $\Delta \log d_{\mathrm{var}} = -0.105$, $-\Delta \log(1+\chi) = -0.271$, and $\Delta \log d_{\mathrm{PR}} = -0.376$ (maximum seedwise reconstruction error 8.9e-16). The off-diagonal term contributes 72.0% of the absolute mean log displacement (paired-seed bootstrap 95% CI [67.9%, 76.0%]). This coordinate-dependent algebraic decomposition carries no causal mediation interpretation.

Temperature calibration absorbs part of the archived baseline NLL contrast. For the linear observer, the uncalibrated and calibrated effects are 0.556 and 0.120; for fixed RFF they are 1.015 and 0.149. Thus NLL endpoint magnitude is partly calibration-mediated and remains observer-contract dependent.

## H Predictive and minimal-mechanism follow-up

We sampled source-recording groups at 10, 20, 40, 70, and 100% budgets, budgeted training and validation labels together, retained the full unlabeled training pool only for scaling, and repeated subset selection five times per model seed. Table 12 reports the realized labeled counts and high-minus-low held-out NLL contrasts. The registered interaction is the per-seed slope of ΔNLL against

Table 12: Five-budget natural-sound learning-curve assay. Counts are median [range] across seed–draw subsets; ΔNLL is high minus low with a paired-seed 95% CI.

| Budget | Train labels | Validation labels | ΔNLL |
|---|---|---|---|
| 10% | 632 [373,850] | 50 [39,68] | -0.0032 [-0.0103,+0.0039] |
| 20% | 776 [541,1047] | 80 [66,95] | -0.0046 [-0.0144,+0.0041] |
| 40% | 1135 [765,1371] | 140 [119,163] | +0.0016 [-0.0094,+0.0123] |
| 70% | 1564 [1317,1681] | 235 [215,248] | +0.0048 [-0.0064,+0.0175] |
| 100% | 1841 [1841,1841] | 314 [314,314] | +0.0041 [-0.0080,+0.0176] |

log budget. It is unsupported (mean slope 0.0042, 95% CI [-0.0023, 0.0114], exact $p = 0.301$). The natural-sound cache therefore does not support a sample-efficiency claim.

The minimal assay used 12 paired seeds, 400 inputs of 120 steps, an $8 \to 24$ random projection, four temporal bins, and five Gaussian-noise repeats (SD 0.20). Its common state update was $x_t = \alpha x_{t-1} + (1-\alpha)\tanh(u_t W)$; event models reported $\mathbf{1}[x_t \geq \theta]$, with optional zero reset, and $\theta$ was calibrated to a 0.10 clean event rate in each condition. Continuous states and event counts are therefore distinct measurement coordinates.

**Minimal-system coordinate comparison**

| Model | Reported feature | Reset |
|---|---|---|
| Continuous leakage | binned state | none |
| Threshold, no reset | binned events | none |
| Threshold and reset | binned events | zero after event |

This 24-unit nonrecurrent synthetic system isolates transformations under a shared synthetic input and differs from the auditory continuous reservoir in Appendix E.1. Continuous leakage reduced $D$ without a detectable reliability effect ($\Delta D = -0.0171$; $\Delta R = 0.00003$). Thresholding without reset produced $\Delta D = -0.0153$ and $\Delta R = 0.0047$. Adding reset reversed both directions ($\Delta D = 0.0118$; $\Delta R = -0.0093$). These synthetic coordinates localize transformation sensitivity and leave the full auditory-SNN mechanism unresolved.

The counterfactual assay used independent binary 2-Hz and 18-Hz labels, 600/150/300 train/validation/test samples, 96 steps, white/fast-colored/slow-colored noise, five label budgets, and validation-selected ridge probes. The preregistered directional prediction was lower high-minus-low error for the slow task and higher error for the fast task. The counterfactual slow task was at floor (mean error 0.0010; 92.5% zero-error cells), so its original sign-only gate is superseded. The fast task contradicted the predicted cost: high-minus-low error was -0.0144 (95% CI [-0.0185, -0.0096], exact $p = 0.0009766$). Accordingly, these assays reject a universal symmetric slow-benefit/fast-cost account.

# I Evidence beyond the primary FSD50K assay

## I.1 Which analyses motivated the confirmatory endpoints?

A $3 \times 3$ native-timescale screen and leave-one-family-out analyses motivate the matched endpoint intervention. Confirmatory endpoint inference uses a separate evidence pool.

## I.2 Does the direction transfer beyond FSD50K?

A TinySOL instrument transfer preserves the global $D$/$R$ directions and is analyzed separately from FSD50K inference.

## I.3 Does reservoir processing improve external neural alignment?

NAT4 neural recordings provide a negative architecture-level stress test: the reservoir does not improve auditory-cortical RSA over the auditory front end. A small cross-validated A1 within-reservoir profile preference provides only a functional anchor, with no membrane-fit or species-validation interpretation.

## I.4 Which exploratory studies are excluded from confirmatory inference?

Connectivity-motif, learning-trajectory, and local-plasticity studies remain exploratory or technically invalid where their registered gates fail. They serve only to define the scope of the primary claim.

The strongest supported claim remains conditional on the tested auditory task, measurement contracts, and sampled timescale range. The same $D \downarrow / R \uparrow$ direction appears after recurrent drive is disabled; the separately activity-calibrated continuous leaky control under its own policy reproduces the reliability effect and shows much weaker dimensional compression. This bounded architecture–compensation stress test supports no architecture-general law and leaves the role of recurrence unresolved. The study establishes no recognition advantage, biological species mapping, connectomic mechanism, or continuous Pareto front.