STATE-GATED LEARNING SHAPES THE GEOMETRY OF COOPERATION AND COMPETITION IN RECURRENT NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning reshapes how units in a recurrent network cooperate and compete. When learning conditions change, can interactions formed earlier remain functionally relevant under the same present conditions? We address this question with State-Gated Explaining-Away Plasticity (SG-EAP), a normative recurrent model in which a supplied state changes activity regularization and inhibitory gain during reconstruction learning. Different learning states bias connectivity among similarly driven units toward cooperative or competitive geometry. These biases remain detectable during adaptation even when networks with different histories share the same final learning condition and evaluation inputs. We then intervene on the retained geometry within the same matched-history pair. Under a common low-gate condition, implanting the similarity-aligned component from the previously high-gate network improves classification with missing features, whereas replacing that component with the low-history value reduces it. Each edited representation is evaluated with a linear readout refitted on complete responses. The reciprocal effect exceeds all 16 tested equal-norm orthogonal exchanges in each of ten independent input batches. It explains part of the history-dependent performance difference and depends on the recurrent background. These results identify an interpretable structural trace through which earlier learning can partially shape computation under feature loss during ongoing adaptation.

1 INTRODUCTION

Recurrent systems must adapt to present conditions while preserving useful effects of earlier learning. Context changes activity immediately and can also act through plasticity to reshape later interactions. These effects are easy to confuse: if two networks are measured in different states, a different interaction map may simply reflect the current operating point. The sharper question is whether earlier learning remains visible after the current learning condition and evaluation input have been matched.

Cortical perturbation maps make this distinction concrete. Influence mapping has revealed feature-specific competition and local amplification motifs in V1 (Chettih & Harvey, 2019), while later experimental and modeling work shows that their balance depends on sensory drive and circuit regime (Russell et al., 2024; Kong et al., 2025). Fixed-weight circuit models explain how current input and recurrent gain can reshape perturbation responses (Rubin et al., 2015; Sanzeni et al., 2023; Sadeh & Clopath, 2020), while adaptive models allow experience to alter the parameters that generate them. Warm-start and adaptive-memory studies establish that learning history can persist in parameters (Ash & Adams, 2020; Fusi et al., 2005; Benna & Fusi, 2016; Kirkpatrick et al., 2017). We identify an interpretable recurrent component along which this history remains measurable and test its computational contribution under matched present conditions. We match the final learning condition and evaluation probe to measure how earlier exposure biases the relationship between input similarity and recurrent connectivity. We then exchange the retained structural component between the same matched-history networks and fit a clean-only readout to each resulting representation.

We introduce **State-Gated Explaining-Away Plasticity (SG-EAP)**, a normative excitatory–inhibitory recurrent model in which a supplied scalar gate $g \in [0, 1]$ changes the pressures acting during

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.

reconstruction learning. We call $g = 0$ the low gate and $g = 1$ the high gate; the compact figure labels "low state" and "high state" denote these supplied values. We summarize one component of the learned geometry by asking whether recurrent parameters become more positive or more negative as the corresponding inputs become more similar. Positive and negative slopes operationally define cooperative and competitive geometry. We first establish how the supplied gate writes this coordinate, then compare networks that reach the same final learning condition through different histories and are evaluated on the same frozen probe.

The paper follows one evidence chain. Low- and high-gate learning first establish an interpretable cooperation–competition coordinate. Matched final conditions then reveal a signed residue of earlier learning on that coordinate. We then reciprocally exchange only this residue within the common-low history pair: masked-input performance shifts in both directions and more than under matched orthogonal edits, although the coordinate explains only part of the total history difference. Longer trajectories, switching controls, and endpoint and cross-task assays define the scope of this central result.

2 State-gated learning and three experimental questions

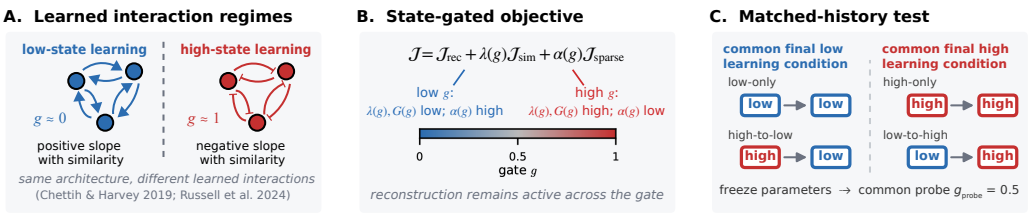


Figure 1: **State-gated learning and the history question. A:** Low-state (low-gate) and high-state (high-gate) learning denote the supplied gate values during learning. At the learned endpoints, the low gate produces a more positive relationship between recurrent interactions and input similarity (cooperative geometry), whereas the high gate produces a more negative relationship (competitive geometry). The cartoon is motivated by feature-specific and contrast-dependent cortical interaction maps (Chettih & Harvey, 2019; Russell et al., 2024; Kong et al., 2025). **B:** Raising g strengthens the similarity-weighted activity penalty and inhibitory gain while reducing the sparsity coefficient; reconstruction remains active throughout. **C:** Different exposure paths enter the same final learning condition; parameters are then frozen and all cohorts are evaluated on matched inputs at the common $g_{\text{probe}} = 0.5$. The directional switching assay is separate.

2.1 From input reconstruction to a state-dependent learning condition

Each sample is supplied directly to 80 excitatory and 20 inhibitory units. A noisy input drives 20 recurrent activity updates, after which the resulting activity is used to reconstruct the corresponding clean input. One slower parameter update differentiates through this recurrent computation. We denote the finite-depth recurrent output after the twentieth implemented step by $\mathbf{r}^*$. Classification labels are reserved for the downstream readout assays and play no role in recurrent learning.

The scalar state $g \in [0, 1]$ is supplied externally. During learning it changes the objective

$$\mathcal{J} = \mathcal{J}_{\text{rec}} + \lambda(g_{\text{learn}})\mathcal{J}_{\text{sim}} + \alpha(g_{\text{learn}})\mathcal{J}_{\text{sparse}}. \tag{1}$$

The implemented state dependence is

$$\lambda(g) = 0.01 + 5g, \qquad \alpha(g) = 0.05 + 0.1(1 - g), \qquad G(g) = 0.2 + g, \tag{2}$$

where G is the fast inhibitory-gain factor. Here $\mathcal{J}_{\text{rec}}$ reconstructs clean inputs from noisy activity, $\mathcal{J}_{\text{sparse}}$ discourages dense excitatory activity, and $\mathcal{J}_{\text{sim}} = (2N_E)^{-1}\langle \mathbf{r}_E^\top S \mathbf{r}_E \rangle$ penalizes activity along directions defined by the input-correlation matrix S. The matrix S retains its unit diagonal, so this term penalizes activity power as well as off-diagonal coactivation. Moving from low to high g therefore strengthens the similarity-weighted activity penalty, decreases the sparsity coefficient, and

increases fast inhibitory gain. In the main experiments these simultaneous changes define one *coupled learning condition*, and the primary comparisons estimate their joint effect. Later ablations assess component-wise sufficiency within this implementation. SG-EAP is a biology-inspired normative framework whose local synaptic implementation remains unspecified. Full dynamics, optimization details, and sign conventions are in Appendix A.

2.2 An interpretable coordinate of recurrent geometry

The structural readout m_W asks a simple question: as the input profiles driving two excitatory units become more similar, does the learned recurrent parameter between those units become more positive or more negative? Formally, m_W is the regression slope between the off-diagonal entries of W_{EE} and the corresponding entries of S. We call a positive slope cooperative and a negative slope competitive. These labels describe a population-level trend; individual entries can have either sign. This scalar captures one interpretable symmetric projection of directed recurrence. Antisymmetric components of W_{EE} are invisible to the symmetric matrix S, and the magnitude of m_W reflects both alignment with S and overall parameter scale.

The E/I labels name model compartments, and the unconstrained W_{EE} represents an effective parameter block. We also measure a functional slope m_K from a finite-difference perturbation kernel of the full frozen network. Structural m_W is computed directly from learned parameters; functional m_K depends on the network's response under the evaluation probe. Their endpoint sign agreement is an empirical result.

We distinguish the state used while parameters learn, g_{learn}, from the state used to evaluate a frozen network, g_{probe}. The sequence of learning states defines the history. Once parameters are frozen, m_W no longer depends on the probe; matching the probe remains essential for comparing m_K and task readouts.

2.3 Three tests of what the geometry records and supports

Writing the geometry. Endpoint networks learn for 600 updates at low ($g_{\mathrm{learn}} = 0$) or high ($g_{\mathrm{learn}} = 1$) state. We then freeze them and measure their structural and functional slopes under the common $g_{\mathrm{probe}} = 0.5$.

Testing learning history. We compare low-only with high-to-low networks under a common final low condition, and high-only with low-to-high networks under a common final high condition. Within each pair, the final 600-update phase has the same learning condition, duration, noise stream, and reinitialized optimizer. At selected checkpoints we freeze both networks and apply the same probe. A remaining difference in m_W therefore measures a structural trace of the earlier learning state during this common final phase.

Testing a computational contribution of retained history. In the common-low history pair, we reciprocally replace only the similarity-aligned coefficient of W_{EE} by its paired value. For each edited representation, we fit a new linear readout to complete responses and hold it fixed while input features are masked. This tests whether the measured history coordinate contributes to the computation supported by that representation. Separate endpoint and task-alignment experiments ask how broadly the same coordinate matters.

An auxiliary rewriting assay switches the learning state in opposite directions from the two endpoints and measures when m_W first crosses zero. Full schedules, optimizer controls, generators, estimators, and the hierarchy of independent input batches and nested parameter initializations are given in Appendices A, B, and E.

3 Learning state writes cooperative or competitive geometry

We first ask whether the supplied learning gate can write the interaction–similarity coordinate defined above. After learning, low-gate networks have positive m_W and high-gate networks have negative m_W in every independent input batch (Fig. 2). The corresponding finite-difference slope m_K changes sign in the same direction when both endpoints are frozen and evaluated under the common probe.

Gate-dependent learning therefore changes both the parameter-level geometry and the direction of influence measured from the full recurrent response.

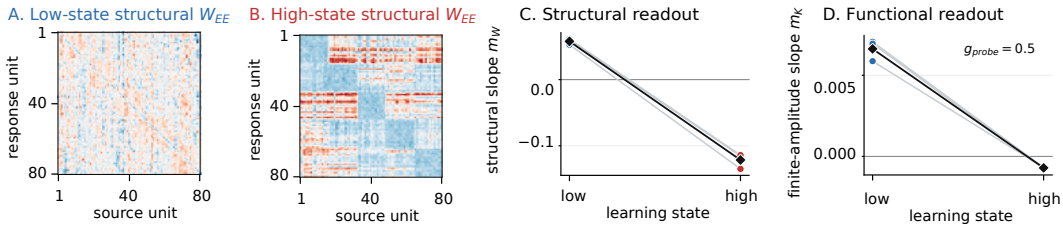


Figure 2: **Low- and high-gate learning write opposite recurrent geometries. A,B:** Representative structural W_{EE} matrices after low- and high-gate learning. Red and blue denote positive and negative entries; each matrix uses its own symmetric display range, making color intensity specific to that panel. **C,D:** Structural and finite-amplitude functional interaction–correlation slopes from the original three independent input batches after all networks are frozen and tested at the common $g_{\mathrm{probe}} = 0.5$. Each colored point is an input-batch mean over three nested initializations; diamonds are grand means. Positive and negative slopes are labeled cooperative and competitive, respectively; individual matrix entries can have either sign. Representative functional matrices evaluated at the two endpoint probes are shown separately in Appendix Fig. 6.

Increasing the activity-penalty coefficient can produce competitive endpoint geometry. A paired factor control separated the three quantities changed by the gate. Starting from shared initial parameters, training with the high-gate coefficient of $\mathcal{J}_{\mathrm{sim}}$ while keeping sparsity pressure and inhibitory gain at their low-gate values produced negative m_W in every tested batch; the sparsity-only and gain-only changes retained positive m_W. This establishes conditional sufficiency with the native S. Necessity and independent factor effects remain untested (Appendix C.6).

The full endpoint reversal also occurs when S is replaced by its diagonal during training, establishing that its off-diagonal entries are unnecessary for this result. In that setting, $\mathcal{J}_{\mathrm{sim}}$ retains its activity-power component while the cross-feature coactivation penalty is removed. Together, the controls show that stronger activity-power pressure can produce reversal within the remaining recurrent reconstruction system with correlated inputs and the other coupled gate changes. The sufficiency of activity power in isolation remains untested (Appendix C.5). The endpoint distinction also holds under held-out and cluster-based geometry readouts and across correlated, Gabor, and natural-input ensembles (Appendix B).

4 Earlier learning leaves a signed trace under the same current condition

4.1 Earlier state remains visible after a common final phase

Measuring history requires a common final learning condition. We therefore created four paths that enter a common low or high final learning phase (Fig. 3A). Within each final condition, paired networks differ only in the state encountered earlier; their final 600 updates use the same learning condition, duration, noise stream, and reinitialized optimizer, followed by the same frozen evaluation probe.

The earlier state remained measurable in structural geometry throughout this window (Fig. 3B,C). After 600 common-final updates, high-to-low networks were more competitive than low-only networks ($\Delta m_W = -0.0116$, 95% CI $[-0.0141, -0.0091]$), whereas low-to-high networks were more cooperative than high-only networks ($\Delta m_W = +0.280$, 95% CI $[0.268, 0.291]$). Both signed contrasts appeared in all ten independent input batches.

The low- and high-final cohorts occupy different absolute ranges of m_W, which restricts inference to the direction within each matched pair. Prior high-gate learning pulls the common low-final path toward competition, and prior low-gate learning pulls the common high-final path toward cooperation. Other frozen readouts retain no aligned difference across both final conditions. We therefore test a

computational contribution in the common-low pair, where both the structural and masked-readout contrasts are reproducible.

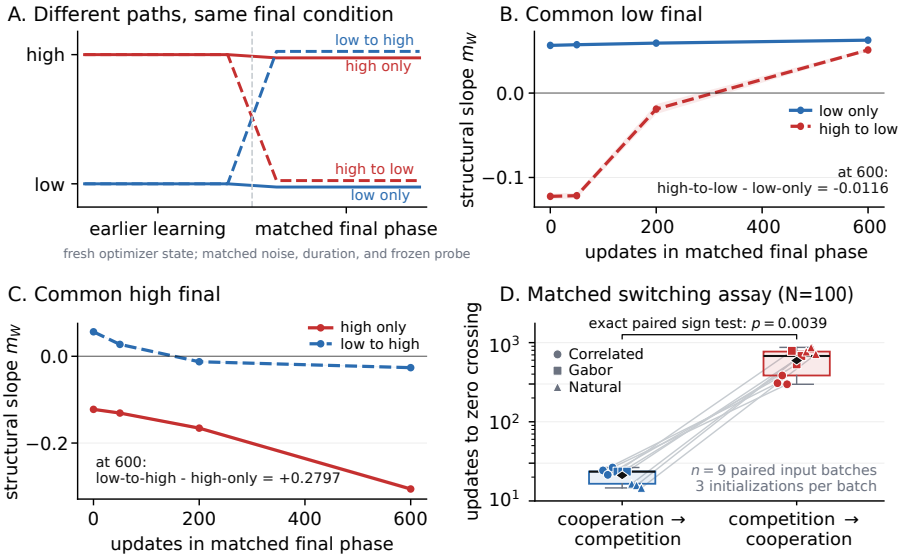


Figure 3: **Earlier learning leaves a signed structural trace under the same current condition. A:** Four exposure paths enter a common low or high final learning condition. **B,C:** Structural geometry during the matched final phases; curves are means ± SEM over ten independent input batches, with three paired initializations nested within each batch. Within each final condition, the switched path remains biased toward its earlier state. **D:** A separate auxiliary switching assay measures zero-crossing times after starting from the two endpoints. Points are nine independent input-batch means, gray lines pair directions, and black diamonds denote means; each batch averages three nested initializations. Boxes show the median and interquartile range, with whiskers extending to $1.5\times$ the interquartile range. Cooperation-to-competition crossed first in 9/9 paired batches (two-sided exact paired sign test, $p = 0.0039$) and all 27 individual runs.

4.2 The retained coordinate partially contributes to common-low computation

The common-low pair lets us ask whether the measured structural trace merely accompanies, or partly contributes to, a functional difference between histories. We use a five-class task in which the label identifies the largest latent source driving the five correlated input groups. At update 600, the high-to-low (HL) representation exceeded the low-only (LL) representation by 3.74 percentage points under 50% feature masking (95% CI $[2.83, 4.66]$; Fig. 4).

Let $\widetilde{S}$ be the centered, zero-diagonal similarity template and write $W_{EE} = a\widetilde{S} + W_{\perp}$, where $a = m_W$ and $W_{\perp}$ is orthogonal to $\widetilde{S}$. For paired LL and HL checkpoints, we exchange only this coefficient:

$$W_{LL\leftarrow HL} = W_{LL} + (a_{HL} - a_{LL})\widetilde{S}, \qquad W_{HL\leftarrow LL} = W_{HL} + (a_{LL} - a_{HL})\widetilde{S}. \tag{3}$$

The first edit implants the HL history coordinate into LL; the second replaces the HL coordinate by its LL value. No edited recurrent network was retrained. Every other parameter and the orthogonal remainder $W_{\perp}$ were held fixed.

Each edited representation received a new linear readout fitted only on complete responses and held fixed for masked evaluation. Implanting the retained coefficient into LL improved masked accuracy by 1.18 points ($[0.72, 1.64]$); replacing it in HL reduced masked accuracy by 0.72 points ($[0.42, 1.01]$). Both effects had the predicted sign in all ten independent input batches. Averaging the reciprocal changes gives a symmetric coordinate effect of 0.95 points ($[0.61, 1.29]$). Its effect on complete-input accuracy was 0.06 points ($[-0.07, 0.18]$), so the masking-specific contribution was 0.89 points ($[0.62, 1.16]$).

The aligned exchange exceeded the mean of 16 equal-Frobenius-norm reciprocal edits orthogonal to the similarity template by 0.97 points ($[0.68, 1.26]$) and ranked first among the 17 directions in all ten batches. In the symmetric four-cell decomposition, the complementary background term was 2.80 points ($[2.07, 3.53]$). This term averages across the two edited backgrounds and differs from the residual of either single-direction edit. The implant effect exceeded the erase effect by 0.46 points ($[0.09, 0.83]$), showing that the contribution depends on the recurrent background. The retained coordinate therefore explains a specific portion of the history difference.

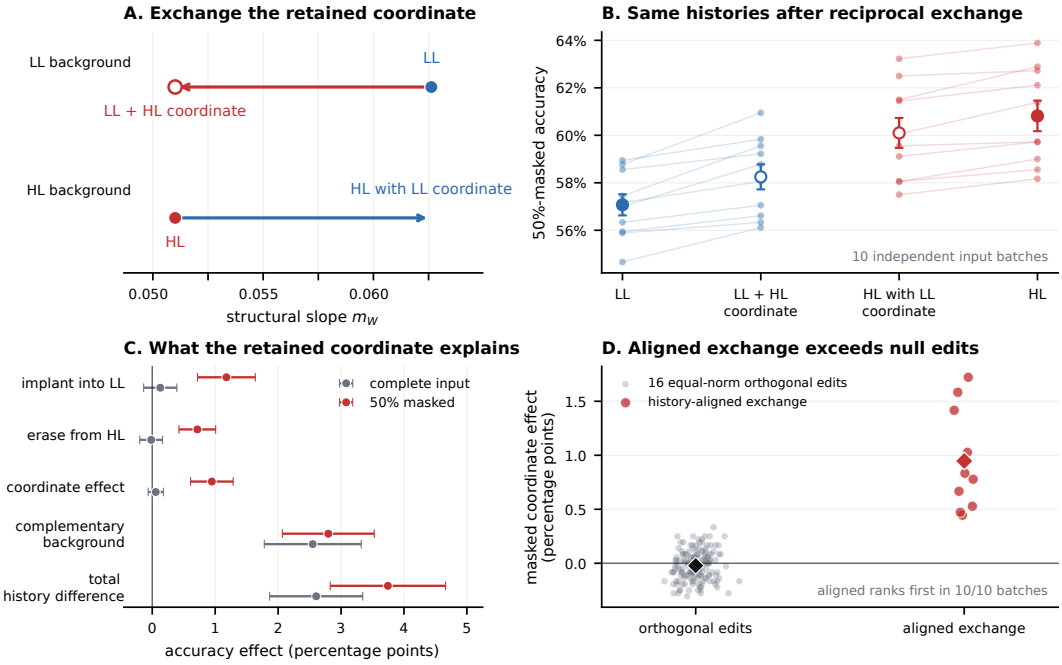


Figure 4: **Reciprocal exchange links retained history geometry to computation. A:** At update 600 of the common-low phase, the similarity-aligned coefficient is exchanged between low-only (LL) and high-to-low (HL) networks while the orthogonal recurrent background is kept fixed. **B:** Masked-input accuracy for the four resulting history–coordinate combinations after a separate clean-only readout is fitted to each representation. Points are ten independent input-batch means, with three paired initializations nested within batch. **C:** Across-batch means and 95% confidence intervals for paired effects on complete and 50%-masked inputs. The symmetric coordinate effect averages the reciprocal implant and erase changes; the complementary background term completes the four-cell identity and differs from either single-direction residual. **D:** Masked coordinate effects for the aligned exchange and 16 equal-norm orthogonal reciprocal edits. Small points are batch-level effects and diamonds are means; the aligned direction ranked first in 10/10 batches.

4.3 The structural trace continues to evolve during extended adaptation

The trajectories were still evolving at update 600, so the common-final result concerns history during ongoing adaptation. We followed the same four paths for 4,800 updates to determine how that trace changed. At equal update counts, both contrasts kept their original directions, but the low-final gap nearly disappeared while the high-final gap expanded. Matching checkpoints by one scalar objective preserved the high-final contrast but reversed the low-final contrast. No branch satisfied all three operational stationarity criteria within the tested horizon, leaving asymptotic convergence unresolved. These runs show that the structural difference depends on both exposure history and the point along the subsequent optimization trajectory (Appendix C.12).

4.4 AN AUXILIARY ASSAY BOUNDS DIRECTIONAL REWRITING

A separate endpoint-switch assay used matched 20-update ramps and observation windows. Cooperation-to-competition crossed zero first in all 27 $N = 100$ runs and all nine independent input-batch means (two-sided exact paired sign test, $p = 0.0039$; Fig. 3D). This ordering is protocol dependent: clearing AdamW's stored state greatly reduced its magnitude, the tested $N = 1000$ calibration yielded a different censored order, and a state-gated Oja-style rule using a direct update recovered the qualitative effect. The assay therefore characterizes rewriting under the reference protocol (Appendices E and G).

5 FUNCTIONAL SCOPE BEYOND THE MATCHED-HISTORY PAIR

The history-specific exchange provides an interventional link within one matched pair. We next ask whether the same structural coordinate matters beyond that pair. In a separate endpoint experiment, the 80 excitatory inputs form five correlated groups and the label identifies the largest latent source. A readout fitted on complete recurrent responses is held fixed while input features are masked. This controlled test measures robustness in the model's native synthetic setting.

At 50% masking, the high-gate endpoint lost 11.81 percentage points less accuracy than the low-gate endpoint. Editing the low endpoint along the full low-to-high similarity coordinate increased masked accuracy by 10.72 points (95% CI $[10.03, 11.42]$) after clean-only readout refitting, while clean accuracy changed by -0.14 points ($[-0.53, 0.25]$). The reduction in clean-to-masked loss exceeded the mean of 16 equal-norm orthogonal edits by 10.14 points ($[9.46, 10.83]$), and the aligned edit ranked first in every batch (Appendix Fig. 8). The endpoint and history-pair effect sizes describe distinct interventions because their coordinate changes and recurrent backgrounds differ.

5.1 CROSS-TASK ALIGNMENT SWEEP

The native five-group task is aligned with the same covariance structure used to define S. We therefore measured how the edit's robustness benefit depends on task alignment. We kept the trained endpoints, recurrent edit, corruption, and readout protocol fixed while rotating a binary task direction between the leading five-dimensional eigenspace of S and its orthogonal complement (Appendix Fig. 10A). The alignment variable q is the squared projection of the unit task vector onto the leading subspace. Task-score variance and difficulty vary independently across conditions.

The edit's effect depended on task direction. The prespecified slope across alignment levels was positive in all ten input batches (mean 0.0204 accuracy-loss units per unit q, 95% CI $[0.0151, 0.0256]$), and the benefit at $q = 1$ exceeded that at $q = 0$ by 2.59 percentage points ($[2.07, 3.12]$; Appendix Fig. 10). At every sampled q, both clean and masked accuracy rose; most of the increase occurred between $q = 0$ and $q = 0.25$, after which it plateaued. The result establishes task selectivity between an orthogonal direction and tasks containing an aligned component. Its nonlinear sampled curve leaves the form of the alignment law open.

Holding the original low-gate readout fixed after the endpoint edit lowers both clean and masked accuracy. Fixed branch-specific decoders likewise reverse the positive reciprocal closure in the history-pair exchange: implanting the HL coefficient lowers LL performance, whereas replacing the coefficient in HL raises performance. The positive edit result therefore measures what a clean-only refitted readout can recover from each representation. Additional controls retain the endpoint gap after simple input filling, reverse its direction in a fixed nonrecurrent spectral reference, and preserve it after diagonal-only training (Appendices C.3, C.4, and C.5).

6 RELATED WORK

Context during inference and during learning. Fixed-weight circuit models and conditional-computation methods alter activity or effective parameters as a function of current context (Rubin et al., 2015; Sanzeni et al., 2023; Ha et al., 2017; Perez et al., 2018). Multi-task and multi-objective methods share parameters across objectives (Caruana, 1997; Sener & Koltun, 2018). SG-EAP uses a supplied learning condition and then holds the present learning condition and probe fixed, isolating the contribution of earlier plasticity from the acute contextual response.

Learned recurrent structure and adaptive memory. Warm starts can alter generalization even under stationary data (Ash & Adams, 2020), and consolidation or continual-learning models allow history to reside in parameters or slower variables (Fusi et al., 2005; Benna & Fusi, 2016; Kirkpatrick et al., 2017). Oja learning, sparse and efficient coding, anti-Hebbian decorrelation, and similarity matching likewise show how input statistics can reshape representations (Oja, 1982; Barlow, 1961; Földiák, 1990; Olshausen & Field, 1996; Pehlevan & Chklovskii, 2019). Our contribution is to match the current learning condition and probe while following a fixed, interpretable component of recurrent structure—how learned parameters covary with input similarity—and then exchange that component within the same matched-history pair. The state-gated Oja-style comparator provides a second realization of the qualitative history effect. SG-EAP supplies an explicit normative account of one shared adaptive phenomenon.

Robustness to incomplete inputs. Associative memories and denoising autoencoders recover or represent structure from incomplete observations (Hopfield, 1982; Ramsauer et al., 2021; Vincent et al., 2010). Our endpoint assay asks a narrower question: how accurately a readout fitted on complete recurrent responses classifies inputs after features are masked, and whether a targeted recurrent edit changes that accuracy.

7 Discussion

This study follows one interpretable recurrent coordinate from learning to history to intervention. Low- and high-gate learning write opposite endpoint geometries. After a common final learning phase, earlier high-gate exposure biases the low-final path toward competition, whereas earlier low-gate exposure biases the high-final path toward cooperation. In the common-low pair, reciprocally exchanging only that retained coordinate shifts masked-input accuracy in both directions and more than equal-norm orthogonal edits. Together, these results establish a signed structural trace of earlier learning and a partial interventional contribution to later computation.

The history result captures ongoing adaptation over the tested window. The longer paths remain in motion, and matching one scalar objective changes the low-final directional contrast. Asymptotic convergence therefore remains open. The auxiliary switching assay has a similarly defined scope: its zero-crossing order depends on optimizer state, network-size calibration, endpoint position, and protocol, and a state-gated Oja-style rule reproduces the qualitative effect. The matched-current structural comparison and reciprocal edit provide the central evidence.

The functional result also has a precise scope. The exchanged coordinate explains a minority of the common-low history difference, and its effect depends on the recurrent background. Positive reciprocal closure appears after fitting a clean-only readout to each edited representation; fixed branch-specific decoders reverse this pattern. The intervention therefore measures what each edited representation can support after readout learning. Immediate improvement to a deployed decoder and complete mediation of history remain outside its scope.

Endpoint and cross-task experiments extend the result within the tested setting. High-gate representations transfer better from complete to masked inputs in the grouped-covariance task, and the full endpoint edit changes the robustness available to a refitted readout more than the tested equal-norm alternatives. The benefit is smaller for an orthogonal task than for tasks containing a component in the leading input-similarity subspace. These experiments cover one synthetic covariance family; benchmark-level robustness and a universal alignment law remain open.

The cortical data in Appendix H motivate the observable by showing that V1 interaction–similarity slopes change with current contrast. The dataset contains no controlled histories or longitudinal recovery measurements. A prospective experiment would compare low-only with high-to-low exposure after a common final low-learning interval, and high-only with low-to-high exposure after a common final high-learning interval, then measure causal maps under the same present probe.

The evidence is thus deliberately bounded: bidirectional structural history is measured during a finite common-final window, functional exchange is established only for the common-low pair, and the switching result is limited to the matched $N = 100$ protocol. SG-EAP also uses an externally supplied gate, simplified inhibition, an unconstrained within-excitatory-state parameter block, and non-biological optimization. Within those boundaries, it provides a concrete way to identify what

prior learning leaves in recurrent geometry and to test whether that retained component contributes to subsequent computation.

REPRODUCIBILITY STATEMENT

The appendix specifies the architecture, objective, generators, metrics, statistical units, controls, and adaptation protocols. The supplementary code-and-results archive provides frozen protocols, simulation and figure-generation code, and saved run-level JSON/CSV summaries. For the history-specific exchange, the released entry point regenerates the omitted recurrent checkpoints and per-example evaluation assets, including masks, fitted readouts, null edits, and cell-level intermediates. The compact upload omits these large reproducible assets. The README maps figures to results and reconciles code and manuscript notation. Development runs are separate and excluded from confirmation statistics.

AI USE STATEMENT

OpenAI GPT models and Codex were used for code review and debugging, analysis auditing, and language editing. This assistance also included hypothesis refinement, feedback on experimental protocols, code implementation, and review of result interpretation. The authors are responsible for the scientific questions, experimental execution, verification of the results, and the final content.

REFERENCES

Jordan Ash and Ryan P. Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3883–3894, 2020.

Horace B. Barlow. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, pp. 217–234. MIT Press, 1961.

Marcus K. Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, 2016. doi: 10.1038/nn.4401.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. doi: 10.1023/A:1007379606734.

Selmaan N. Chettih and Christopher D. Harvey. Single-neuron perturbations reveal feature-specific competition in V1. *Nature*, 567(7748):334–340, 2019. doi: 10.1038/s41586-019-0997-6.

Peter Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological cybernetics*, 64(2):165–170, 1990. doi: 10.1007/BF02331346.

Stefano Fusi, Patrick J. Drew, and L. F. Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005. doi: 10.1016/j.neuron.2005.02.001.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Deyue Kong, Joe Barreto, Lorenzo Butti, Matthias Kaschube, and Benjamin Scholl. Principles of cortical interactions in modular recurrent networks. bioRxiv preprint, December 2025. URL https://doi.org/10.64898/2025.12.19.695176. Version 1 posted 2025-12-21; version 2 posted 2026-01-29.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. doi: 10.1038/381607a0.

Cengiz Pehlevan and Dmitri B Chklovskii. Neuroscience-inspired online unsupervised learning algorithms: artificial neural networks. *IEEE Signal Processing Magazine*, 36(6):88–96, 2019. doi: 10.1109/MSP.2019.2933846.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Daniel B. Rubin, Stephen D. Van Hooser, and Kenneth D. Miller. The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2): 402–417, 2015. doi: 10.1016/j.neuron.2014.12.026.

Lloyd E. Russell, Mehmet Fişek, Zidan Yang, Lynn Pei Tan, Adam M. Packer, Henry W. P. Dalgleish, Selmaan N. Chettih, Christopher D. Harvey, and Michael Häusser. The influence of cortical activity on perception depends on behavioral state and sensory context. *Nature Communications*, 15(1): 2456, March 2024. ISSN 20411723. doi: 10.1038/s41467-024-46484-5.

Sadra Sadeh and Claudia Clopath. Theory of neuronal perturbome in cortical networks. *Proceedings of the National Academy of Sciences*, 117(43):26966–26976, 2020. doi: 10.1073/pnas.2004568117.

Alessandro Sanzeni, Agostina Palmigiano, Tuan H. Nguyen, Junxiang Luo, Jonathan J. Nassi, John H. Reynolds, Mark H. Histed, Kenneth D. Miller, and Nicolas Brunel. Mechanisms underlying reshuffling of visual responses by optogenetic stimulation in mice and monkeys. *Neuron*, 111(24): 4102–4115.e9, 2023. doi: 10.1016/j.neuron.2023.09.018.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014. doi: 10.7717/peerj.453.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

A Implementation details

A.1 Architecture and optimization

Let $g_{\mathrm{learn}} \in [0, 1]$ denote the supplied state during parameter learning and g_{probe} the state supplied to a frozen network. Finite training with algorithm $\mathcal{A}_T$ gives

$$\theta_T(g_{\mathrm{learn}}) = \mathcal{A}_T(\theta_0, \mathcal{D}; g_{\mathrm{learn}}), \qquad \mathbf{r}^* = F_{\theta_T(g_{\mathrm{learn}})}(\mathbf{u}; g_{\mathrm{probe}}). \tag{4}$$

Here $\mathbf{r}^*$ is the finite-depth terminal activity after the implemented 20-step unroll. All analyses use this output directly, without an equilibrium approximation. The implemented objective is

$$\mathcal{J}(\theta; g_{\text{learn}}) = \mathcal{J}_{\text{rec}} + \lambda(g_{\text{learn}})\mathcal{J}_{\text{sim}} + \alpha(g_{\text{learn}})\mathcal{J}_{\text{sparse}}, \tag{5}$$

with $\lambda(g) = \lambda_e + \gamma_e g$ and $\alpha(g) = \lambda_a + \gamma_a(1-g)$.

The E/I network uses 80 excitatory and 20 inhibitory state variables in canonical experiments. With the post←pre convention, its activity obeys

$$\tau_r \dot{\mathbf{r}}_E = -\mathbf{r}_E + \phi(\mathbf{u}_E + W_{EE}\mathbf{r}_E - G(g)c_{\text{inh}}W_{EI}\mathbf{r}_I + \mathbf{b}_E)\,, \tag{6}$$

$$\tau_r \dot{\mathbf{r}}_I = -\mathbf{r}_I + \phi(\mathbf{u}_I + W_{IE}\mathbf{r}_E - G(g)c_{\text{inh}}W_{II}\mathbf{r}_I + \mathbf{b}_I)\,, \tag{7}$$

where ϕ is ReLU, $G(g) = 0.2 + g$, and $c_{\text{inh}} = 2$. The unconstrained within-excitatory-state block represents effective recurrent parameters. The other recurrent blocks use softplus parameterizations with the signs displayed above. The released implementation names cross-population matrices by pre→post direction, so its two cross-population labels are exchanged relative to the text's post←pre notation.

The implemented forward computation uses $\tau_r = 1$ and explicit Euler steps,

$$\mathbf{r}^{(t+1)} = \mathbf{r}^{(t)} + 0.1\left[-\mathbf{r}^{(t)} + \phi\Big(\mathbf{u} + W_{\text{eff}}(g)\mathbf{r}^{(t)} + \mathbf{b}\Big)\right], \qquad t = 0, \ldots, 19. \tag{8}$$

Every sample starts from $\mathbf{r}^{(0)} = 0$ on every network call; the supplied input is held constant across the 20 steps. Training examples and every finite-difference perturbation are therefore evaluated from a freshly reset activity state.

The expanded objective terms in Eq. 5 are

$$\mathcal{J}_{\text{rec}} = \frac{1}{2N}\left\langle \|\mathbf{r}^* - \mathbf{u}\|^2 \right\rangle, \tag{9}$$

$$\mathcal{J}_{\text{sim}} = \frac{1}{2N_E}\left\langle \mathbf{r}_E^{*\top} S \mathbf{r}_E^* \right\rangle, \tag{10}$$

$$\mathcal{J}_{\text{sparse}} = \frac{1}{N_E}\left\langle \sum_{i \in E} \sqrt{r_{E,i}^* + 10^{-6}} \right\rangle. \tag{11}$$

Angle brackets denote the training-input average. The standard-correlation matrix S retains its unit diagonal, so $\mathcal{J}_{\text{sim}}$ penalizes activity power as well as off-diagonal coactivation. Recurrent matrices receive scaled Kaiming initialization (He et al., 2015), and biases start at zero. Each outer update presents the full batch, adds independent Gaussian noise of SD 0.3, unrolls 20 activity steps with $\Delta t = 0.1$, backpropagates through the unroll, clips the global gradient norm at 1, and makes one AdamW step with zero weight decay (Loshchilov & Hutter, 2019). The learning rate is 10^{-3} for correlated and natural inputs and 2×10^{-3} for Gabor inputs.

The canonical coefficients are $(\lambda_e, \gamma_e) = (0.01, 5.0)$ and $(\lambda_a, \gamma_a) = (0.05, 0.1)$. Inhibitory input has SD 0.05. Endpoint and history experiments use 600 updates per phase. Structural slopes exclude the diagonal.

The operational geometry readouts used in the main text are

$$m_W = \frac{\text{Cov}((W_{EE})_{\text{off}}, S_{\text{off}})}{\text{Var}(S_{\text{off}})}, \tag{12}$$

$$K_{EE,ij}^{(h)}(g_{\text{probe}}) = \mathbb{E}_{\mathbf{u}}\left[\frac{F_{\theta,E,i}(\mathbf{u} + h\mathbf{e}_j; g_{\text{probe}}) - F_{\theta,E,i}(\mathbf{u} - h\mathbf{e}_j; g_{\text{probe}})}{2h}\right], \quad h = 0.05, \tag{13}$$

$$m_K = \frac{\text{Cov}((K_{EE}^{(h)})_{\text{off}}, S_{\text{off}})}{\text{Var}(S_{\text{off}})}. \tag{14}$$

For an unambiguous projection form, define the zero-diagonal template

$$\bar{S}_{\text{off}} = \frac{1}{N_E(N_E-1)}\sum_{i \neq j} S_{ij}, \qquad \widetilde{S}_{ij} = \begin{cases} S_{ij} - \bar{S}_{\text{off}}, & i \neq j, \\ 0, & i = j. \end{cases} \tag{15}$$

Then $m_W = \langle W_{EE}, \widetilde{S}\rangle_F / \|\widetilde{S}\|_F^2$ and $W_{EE} = m_W \widetilde{S} + W_\perp$, with $\langle W_\perp, \widetilde{S}\rangle_F = 0$. Equivalently, $m_W = \rho_{WS}\sigma_{W,\text{off}}/\sigma_{S,\text{off}}$: its magnitude combines alignment and weight scale. Because $\widetilde{S}$ is symmetric, the projection reads the symmetric component of W_{EE}; antisymmetric and other template-orthogonal components lie outside this summary.

We quantify functional influence with this finite-amplitude centered response. It equals the local derivative when the interval remains within one linear region; if a ReLU boundary is crossed, it averages the response over that interval. On the first 128 held-out clean inputs, without added noise, we estimate it as

$$(\widehat{K}_{EE}^{(0.05)})_{ij} = \frac{1}{128}\sum_{n=1}^{128} \frac{F_{\theta,E,i}(\mathbf{u}_n + 0.05\mathbf{e}_j; g_{\text{probe}}) - F_{\theta,E,i}(\mathbf{u}_n - 0.05\mathbf{e}_j; g_{\text{probe}})}{0.10}. \tag{16}$$

Only excitatory inputs are perturbed, and activity is reset to zero for every evaluation. The earlier raw missing-coordinate diagnostic is

$$B_{\text{fid}} = \text{MSE}_{\text{miss}}(\mathbf{u}_{\text{mask}}, \mathbf{u}) - \text{MSE}_{\text{miss}}(F_\theta(\mathbf{u}_{\text{mask}}; g_{\text{probe}}), \mathbf{u}). \tag{17}$$

Positive values mean that the network moves missing coordinates toward their clean values in its native output units. Appendix C.3 documents its dependence on output scaling and excludes it from the evidence for cue-dependent reconstruction.

A.2 Input generators

The correlated construction partitions the 80 excitatory coordinates into five groups of 16. Within each group, a shared Gaussian latent factor and independent Gaussian components are mixed at correlation coefficient 0.9 and rectified. The Gabor input uses filter responses from a bank with varied orientation and phase. Natural inputs use mean-centered and normalized 8×10 patches sampled from the `camera`, `coffee`, and `astronaut` example images distributed with scikit-image (van der Walt et al., 2014). The 20 inhibitory coordinates receive independent low-amplitude Gaussian input. Independent input batches regenerate the samples and the correlation matrix S; denoising noise is newly sampled at each training update.

B Endpoint geometry and readout checks

Figure 5 shows the representative optimization trajectories and the raw weight–similarity construction of the structural slope; the endpoint matrices and aggregate estimates appear only in main Fig. 2.

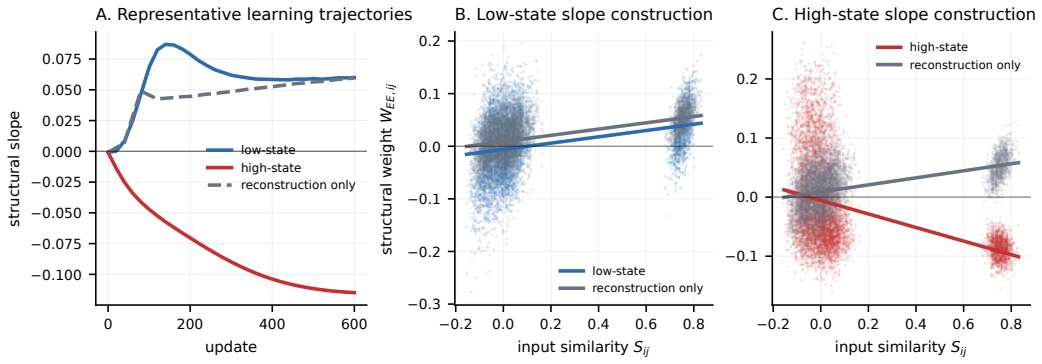


Figure 5: **Representative endpoint-learning diagnostics. A:** Structural interaction slope during low-state, high-state, and reconstruction-only learning. **B,C:** The off-diagonal weight–similarity relation used to construct the representative low- and high-state slopes, each shown with the reconstruction-only baseline. These panels illustrate optimization and slope construction; aggregate endpoint geometry is reported in Fig. 2.

Across fixed-batch five-initialization endpoint ensembles, mean structural slopes $(m_{g=0}, m_{g=1})$ were $(0.059, -0.114)$ for correlated inputs, $(0.253, -0.396)$ for dense Gabor inputs, and $(0.057, -0.131)$

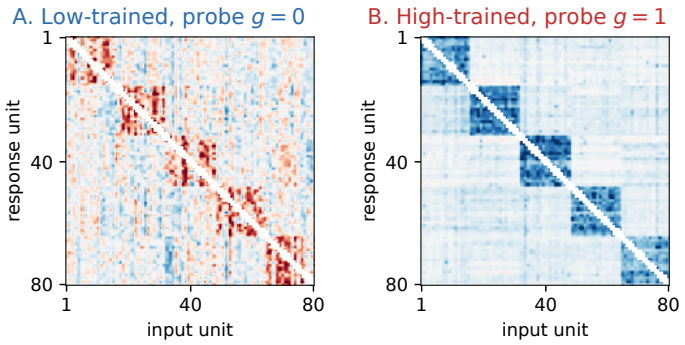


Figure 6: **Representative finite-amplitude functional kernels at the endpoint probes.** The low-trained network is evaluated at $g_{\text{probe}} = 0$ and the high-trained network at $g_{\text{probe}} = 1$. These matrices illustrate response patterns under their respective endpoint probes. Main Fig. 2D reports the separate matched-$g_{\text{probe}} = 0.5$ aggregate.

for dense natural-image inputs; every initialization had the displayed sign. In the representative correlated endpoint, the finite-amplitude functional slope was positive at low gate and negative at high gate. Off-diagonal $K_{EE}^{(0.05)}$ and W_{EE} entries correlated at 0.88 and 0.54, respectively. These empirical checks characterize consistency between parameter and response matrices without equating them.

Three controls address the use of S in both training and evaluation. A held-out S preserved roughly 80–93% of the structural slope. Shuffling the evaluation matrix collapsed the slope. A cluster-based within-minus-across influence statistic, computed independently of regression on S, was positive at the low endpoint and negative at the high endpoint. The paired gate-factor control in Appendix C.6 separately tests which coordinate changes are sufficient from the low-state anchor; other gain-modulated models can switch by different mechanisms.

Approximate symmetry and Gaussian fluctuations provide optional intuition connecting stationary covariance to linear response. Training, the frozen computational probe, and the history tests are defined independently of these assumptions. Direct finite-difference kernels establish the endpoint signs.

C Learning-state, intervention, and history confirmation

C.1 Confirmation protocol

The confirmation protocol was fixed before these results were evaluated. It tested whether learning state writes recurrent geometry and a bounded computational bias under a matched probe, and whether prior learning remains measurable over a fixed common-final window. Independently regenerated correlated-input batches are the inferential units, with paired parameter initializations nested within batch. Before confirmation, development runs set the 600-update history checkpoint and selected the low-trained background for the primary aligned intervention; those runs are excluded from all intervals and sign counts. A later evaluation-only audit rejected raw 50%-mask reconstruction gain as a useful endpoint, and the negative result is reported below.

Confirmation uses input-batch seeds 1302–1304 and parameter seeds 2302–2304, for nine paired cells. Each cell contains six 600-update learning-state endpoints, two five-level aligned interventions, and four history branches. Every frozen evaluation uses the same independently generated 1,200-sample set within a cell and $g_{\text{probe}} = 0.5$. Masks remove 25%, 50%, or 75% of excitatory features; Gaussian corruptions have SD 0.1, 0.3, or 0.5. All confirmation endpoints and probes were finite and non-degenerate.

The readout uses the 80 excitatory responses. On the first 600 complete-response rows, each feature is centered and divided by its population SD; a scale below 10^{-4} is replaced by one. With standardized design Z, five-class one-hot targets Y, and class-frequency intercept $\bar{Y}$, the fitted coefficient matrix

minimizes

$$\|ZB - (Y - \bar{Y})\|_F^2 + \|B\|_F^2. \tag{18}$$

The data-fit term uses an unnormalized sum over samples. The fitted standardizer, B, and intercept are then held fixed while the disjoint final 600 rows are evaluated. Per-edit refitting repeats this complete-response fit separately for each edited representation before its masked test.

For the two headline analyses with only three original inferential units, we froze an independent-batch extension before evaluating it. The original batches 1302–1304 were retained and seven prespecified batches 1305–1311 were added, each with the same three parameter seeds. The extension covered the four-path structural-history assay and the low-endpoint aligned-versus-orthogonal edit. The state sweep, functional-history readouts, and switching experiment retain their original batch counts. Every original batch, initialization, and null direction was preserved. All 21 new batch-by-initialization cells and their edit evaluations passed the prespecified validity checks. Thus the development run selected the metric and checkpoint, the three-batch confirmation fixed the initial tests, later controls were added while preserving those batches, and the seven-batch extension tested the two frozen headline analyses.

For every ten-batch contrast we first average the three paired initialization-level differences within batch, $d_b = 3^{-1}\sum_{s=1}^{3} d_{b,s}$, and report $\bar{d} \pm t_{0.975,9}\,\mathrm{sd}(d_b)/\sqrt{10}$. The interval therefore describes variation across independently regenerated input batches, conditional on the three tested initializations. All batches belong to the same task family.

C.2 Learning-state sweep

Table 1 summarizes batch-level linear trends across the six sampled learning states. These slopes summarize direction across the sampled range, while Fig. 7 shows the actual shape: a sharp change from 0 to 0.2 followed by saturation or partial reversal. The grid is uniform in g and in the absolute value of λ, and its six points leave the response shape within the first interval unresolved. The sweep diagnoses this parameterization. Threshold, phase-transition, and bifurcation interpretations require finer sampling and additional analysis.

Table 1: **Learning-state sweep at common $g_{\mathrm{probe}} = 0.5$.** Slopes are computed across $g_{\mathrm{learn}} \in \{0, .2, .4, .6, .8, 1\}$ within each independent input-batch mean and summarized across three batches. Accuracy is expressed as a proportion.

Readout	Mean slope	95% CI	Batch signs
Structural geometry m_W	-0.1550	$[-0.1790, -0.1311]$	3/3 negative
Functional geometry m_K	-0.00466	$[-0.00589, -0.00344]$	3/3 negative
Raw coordinate-gain diagnostic	-0.00241	$[-0.00372, -0.00111]$	3/3 negative
Masked-test readout accuracy	$+0.1029$	$[0.0790, 0.1268]$	3/3 positive

From $g_{\mathrm{learn}} = 0$ to 1, masked-test readout accuracy changes by $+9.46$, $+15.89$, and $+22.15$ percentage points at 25%, 50%, and 75% masking; every three-batch interval excludes zero. The raw native-output diagnostic changes in the opposite direction, and affine calibration reverses that endpoint ordering under masking. Gaussian corruption produces another ordering. This dependence on output scale and corruption definition excludes the raw coordinate metric from evidence for a fidelity–separation frontier.

C.3 Evaluation-only reconstruction and readout references

We regenerated only the recorded held-out inputs and masks; no network was instantiated or retrained. The first 600 examples were used to fit each reference and the disjoint final 600 examples were used once for evaluation. Recomputed zero-fill errors and the previously saved input-readout cells matched the confirmation file exactly.

For the classification calibration, let $M\mathbf{x}$ be the zero-masked excitatory input and q the prespecified mask fraction. Training-mean fill replaces each missing coordinate by its mean estimated from the complete fitting rows. Known-rate correction evaluates the same readout fitted on complete inputs

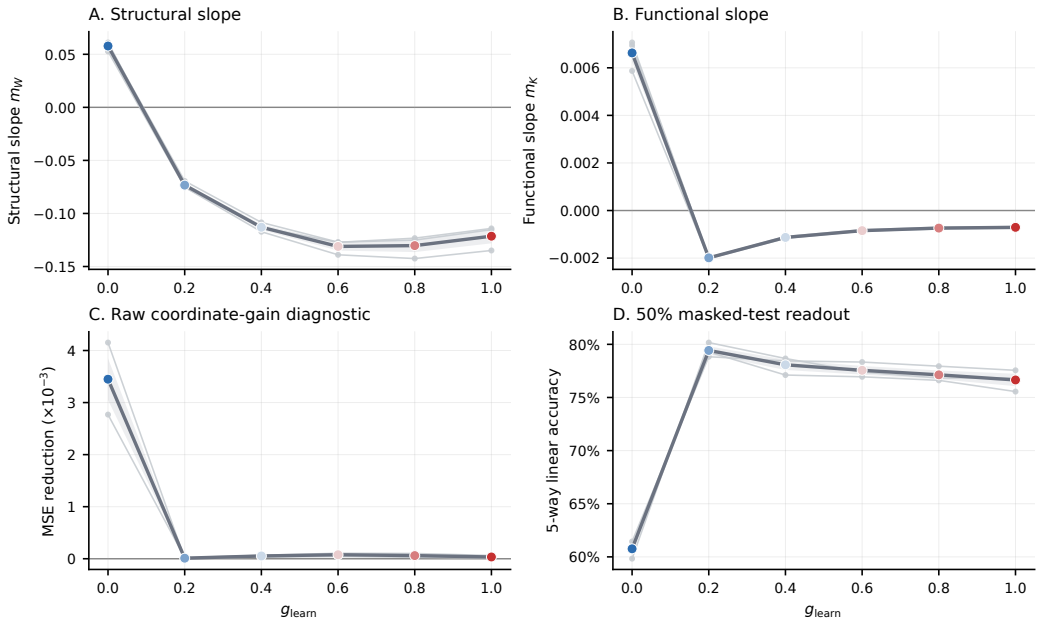


Figure 7: **Six-state diagnostic under a common frozen probe.** Paired networks learn at six supplied values of g_{learn} and are evaluated at $g_{\text{probe}} = 0.5$. **A–D:** Structural geometry, functional geometry, a raw coordinate-reconstruction diagnostic, and 50%-masked readout accuracy change sharply between 0 and 0.2 and then saturate or partially reverse. This first interval also changes the $\mathcal{J}_{\text{sim}}$ coefficient from 0.01 to 1.01. Thin curves are independent input-batch means; thick curves are means over three batches.

on $M\mathbf{x}/(1-q)$. Both corrections use only the complete fitting data and the prespecified mask rate. They remove two simple mean/amplitude shifts and cover a limited set of preprocessing methods.

Table 2: **Input reconstruction references at 50% masking.** Missing-coordinate MSE is summarized over three independent input-batch means. The linear imputer receives the same masked input as the recurrent network but is supervised with clean targets on its fitting split.

Input estimator	Test MSE	Gain over zero fill
Zero fill	0.5017	0
Training-feature mean	0.3437	0.1580
Linear ridge imputation	0.1536	0.3481

For scale, the saved recurrent-network raw gains are 0.00345 after low-state learning and 0.000034 after high-state learning, evaluated over all 1,200 examples. Even the cue-free mean reference is therefore roughly two orders of magnitude larger. This invalidates zero fill as evidence of meaningful cue-dependent reconstruction. The original three-batch confirmation lacks saved endpoint tensors and per-example responses, leaving its missing rows unrecoverable. Later frozen controls regenerated and saved separate endpoint states for their own result sets.

At 50% masking, mean fill explains 10.6% of the 12.89-point high-state advantage over the uncorrected input readout fitted on complete examples; known-rate scaling explains none. The high-state endpoint remains 11.52 points above mean-filled input (95% CI $[9.45, 13.59]$) and 12.91 points above rate-corrected input ($[9.44, 16.38]$). Its 76.65% accuracy is close to the 75.67% input reference fitted on masked examples. These results establish robustness from complete-response fitting to masked evaluation within this assay. Benchmark superiority and intrinsic representation separation remain outside its scope.

Table 3: **Dominant-group accuracy at 50% masking.** Mean fill uses feature means from complete fitting rows; known-rate correction rescales observed coordinates by $1/(1-0.5)$. Both reuse the input readout fitted on complete examples without masked-test labels. The readout fitted on masked inputs is a differently trained reference.

Representation	Decoder fit	Test correction	Clean	Masked
Input	Clean	Zero fill	81.50%	63.76%
Input	Clean	Training-mean fill	81.50%	65.13%
Input	Clean	Known-rate scaling	81.50%	63.74%
Input	50%-masked	Zero fill	87.83%	75.67%
Low-state endpoint	Clean	—	79.61%	60.76%
High-state endpoint	Clean	—	83.69%	76.65%

C.4 Fixed spectral reference

We also tested the fixed, nonlearned map $T_\lambda = (I + \lambda S)^{-1}$ as a deliberately simple spectral reference. For each of 10 independently generated training batches, S was estimated from the same 300 grouped-correlated inputs used by SG-EAP. We applied T_λ to both clean and masked excitatory inputs, fitted the same standardized ridge readout on 600 clean transformed examples, and evaluated it once on 600 disjoint examples. Three evaluation seeds were averaged within each batch. The coefficient grid $\lambda \in \{0, 0.01, 0.03, 0.1, 0.3, 1, 3, 5.01\}$ and all masks were fixed before evaluation; no coefficient was chosen using masked-test labels.

Table 4: **Fixed spectral reference under feature masking.** Entries are percentages summarized over 10 independent input-batch means. The final column is high minus low masked accuracy, comparing $\lambda = 5.01$ with $\lambda = 0.01$; intervals are 95% t intervals across batches.

Mask	$\lambda = 0.01$	$\lambda = 5.01$	High − low
25%	72.30	71.52	-0.78 $[-0.96, -0.61]$
50%	62.33	61.32	-1.01 $[-1.20, -0.81]$
75%	48.19	47.43	-0.77 $[-1.14, -0.39]$

The full prespecified curve declined gradually with stronger shrinkage. All 10 batches had a negative high-minus-low contrast at 50% masking; nine of 10 did so at 75%. This direction is opposite to the SG-EAP endpoint result. For $\lambda \geq 0$ the transform is invertible, so before regularization it preserves the class of affine readouts. With a fixed ridge penalty and feature standardization, however, it changes the effective regularization imposed on the corresponding input-space predictor. It therefore serves as one explicit reference with a different representation and regularization structure. Matched-history residue and the necessity of recurrence remain untested by this comparison.

C.5 Diagonal-only similarity during training

We froze this control before evaluating its confirmation runs. For the same three input batches and three paired parameter initializations used in the native-S confirmation, we trained new low- and high-state endpoints after replacing S only inside the loss by $D = \mathrm{diag}(\mathrm{diag}\, S)$. Initial parameters, inputs, paired training-noise streams, 600-update budget, optimizer, coefficients, gain schedule, sparsity term, and common-probe evaluation were unchanged; no coefficient was retuned. Geometry was still evaluated against the native input similarities. All 18 trained endpoints passed the prespecified finite and endpoint-validity checks.

Under diagonal-only training, the low/high structural slopes against native similarity were $+0.07124/-0.10372$, and the corresponding functional slopes were $+0.01045/-0.001777$. The control rules out a necessary contribution from native off-diagonal entries to either the geometry reversal or the robustness interaction. It leaves the source unresolved because the diagonal activity-power term, training-time gain, sparsity, reconstruction dynamics, and their interactions remain bundled.

For scale, at 50% masking the diagonal-only low endpoint achieved 79.13% clean and 58.46% masked accuracy, whereas the high endpoint achieved 86.00% clean and 79.57% masked accuracy.

Table 5: **Robustness with native or diagonal-only training similarity.** Entries are percentage-point robustness interactions, defined as low-state accuracy loss minus high-state accuracy loss. Intervals use three independent input-batch means; paired parameter initializations are nested within batch.

Masked	Native S	Diagonal D	Native − diagonal	Batch signs
25%	5.39 [4.31, 6.47]	6.24 [4.01, 8.47]	−0.85 [−3.12, 1.42]	3/3 negative
50%	11.81 [8.85, 14.78]	14.24 [11.00, 17.48]	−2.43 [−4.70, −0.15]	3/3 negative
75%	18.07 [17.99, 18.15]	20.57 [16.54, 24.61]	−2.50 [−6.46, 1.46]	3/3 negative

The high endpoint therefore combines a smaller masking loss with higher absolute clean and masked accuracy, excluding a lower clean baseline as the source of the interaction in Table 5.

C.6 Which gate factor is sufficient from the low-state anchor?

We next separated the three coordinates changed jointly by the gate. Starting from the low-state values, we changed only the similarity-weighted activity coefficient λ, only the sparsity coefficient α, only fast inhibitory gain G, or both objective coefficients while retaining low-state gain. Canonical low and full-high endpoints served as references. Three independently generated correlated-input batches were crossed with three paired initializations; within each cell, all six conditions shared initial parameters, inputs, training-noise streams, and the 600-update budget. Table 6 reports means and 95% t intervals over the three input-batch means.

Table 6: **Gate-factor sufficiency from the low-state anchor.** Δm_W is paired against the low reference. Initializations are nested within input batch.

Condition	Endpoint m_W	Δm_W from low	Interpretation
Low reference	+0.05799 [+0.05558, +0.06041]	–	Reference
Full high	−0.11579 [−0.12502, −0.10656]	−0.17379 [−0.18275, −0.16483]	High reference
λ only	−0.08573 [−0.08952, −0.08194]	−0.14372 [−0.14559, −0.14185]	Sign reversal
α only	+0.05457 [+0.05287, +0.05627]	−0.00342 [−0.00464, −0.00220]	Partial shift
G only	+0.05466 [+0.05238, +0.05694]	−0.00333 [−0.00364, −0.00302]	Partial shift
Objective only	−0.09179 [−0.09638, −0.08720]	−0.14978 [−0.15280, −0.14677]	Sign reversal

The λ-only and objective-only endpoints were negative in all three input batches; α-only and G-only endpoints remained positive in all three. Raising the similarity-weighted activity pressure is therefore sufficient for reversal from this anchor. This additive experiment leaves necessity, factor independence, and alternative routes to competitive geometry open. The controls separately establish sufficiency of higher λ with native S and dispensability of native off-diagonal entries under the full coupled-gate change. Sufficiency of diagonal activity pressure in isolation remains untested.

C.7 Similarity-aligned intervention

Let $\widetilde{S}$ be the centered off-diagonal similarity template. We decompose

$$W_{EE} = a\widetilde{S} + W_{\perp}, \qquad a = \frac{\langle (W_{EE})_{\mathrm{off}}, \widetilde{S}_{\mathrm{off}} \rangle}{\|\widetilde{S}_{\mathrm{off}}\|_F^2}. \tag{19}$$

Because $\widetilde{S}$ is centered and normalized to unit regression slope, the coefficient a equals the structural slope m_W used in the main text.

For each paired endpoint we center the off-diagonal similarity template and compute the projection in Eq. 19. Five interpolation fractions replace a between its paired endpoint values while keeping $W_{\perp}$, every other recurrent block, biases, g_{probe}, and evaluation examples fixed. At every interpolation point, the feature standardizer and ridge readout are refitted on clean responses from that edited network and then held fixed for masked evaluation. Table 7 reports linear trends across interpolation fraction. In the low-trained background, m_K and masked-test accuracy move consistently, while clean-test accuracy remains unchanged. In the high-trained background, m_K changes slightly whereas both accuracy intervals include zero. Raw and affine-calibrated coordinate-reconstruction diagnostics

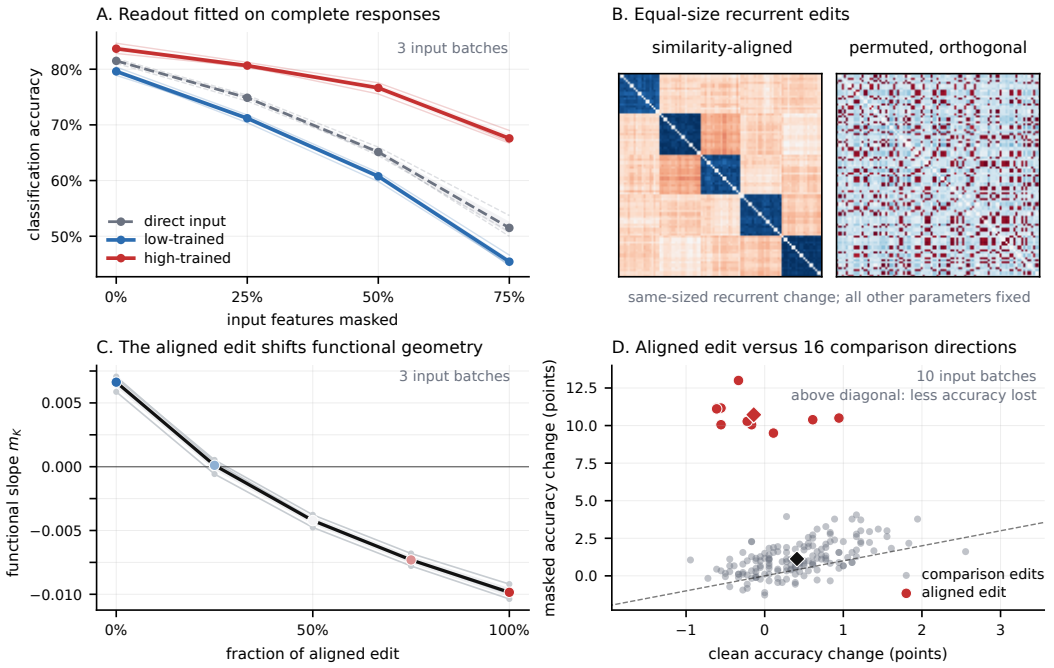


Figure 8: **Computational consequence of the learned endpoint geometry. A:** A readout fitted only on complete responses is tested as input features are masked. Thin lines are the original three input-batch means and thick lines their means. The direct-input reference uses training-mean fill. **B:** The tested similarity-aligned recurrent edit and one representative equal-Frobenius-norm comparison edit. **C:** The aligned edit moves the finite-difference slope toward competition in the original three-batch dose-response experiment. **D:** Across ten independent batches, the refitted readout gives the aligned edit a larger reduction in the accuracy loss from complete to masked inputs than the 16 tested comparison directions. Small points are direction-by-batch means; diamonds are grand means.

follow different curves (Fig. 9). The dose response supports a background-dependent contribution; sufficiency and sole mediation remain untested. The equal-norm comparison below evaluates the aligned edit against a prespecified set of unaligned directions.

Table 7: **Similarity-aligned intervention.** Values are batch-level slopes as the aligned coefficient moves from the cooperative to the competitive endpoint value. Accuracy is a proportion.

Background	Readout	Mean slope	95% CI	Batch signs
Low-trained	m_K	-0.01614	$[-0.01736, -0.01493]$	3/3 negative
	Clean-test accuracy	-0.00126	$[-0.01006, 0.00754]$	1/3 positive
	Masked-test accuracy	$+0.10156$	$[0.07607, 0.12704]$	3/3 positive
High-trained	m_K	-0.000162	$[-0.000185, -0.000139]$	3/3 negative
	Clean-test accuracy	-0.00385	$[-0.01084, 0.00314]$	0/3 positive
	Masked-test accuracy	$+0.00844$	$[-0.00255, 0.01944]$	3/3 positive

C.8 Equal-norm aligned-versus-unaligned comparison

We evaluated 30 low-state endpoints across ten independent input batches: nine endpoints from the original confirmation and 21 from the prespecified extension. The original endpoints were required to reproduce their frozen final loss, m_W, clean accuracy, and 50%-masked accuracy before accepting any edit; the extension used the same validation criteria directly after training. The full similarity-aligned edit was compared with eight prespecified coordinate-permuted templates applied with both signs. Each of the 16 null edits was symmetric, zero-diagonal, projected orthogonal to the similarity

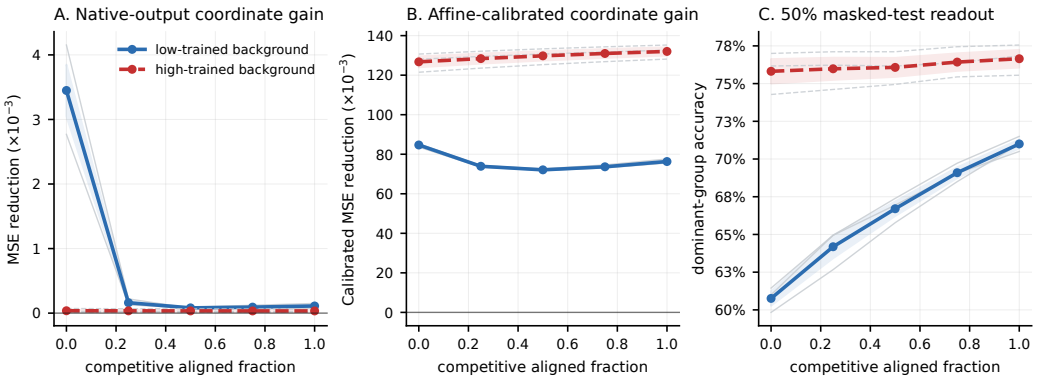


Figure 9: **Coordinate and task diagnostics under the aligned edit. A,B:** Native-output and affine-calibrated coordinate gains follow different trajectories under the same edits, revealing their sensitivity to output scaling. **C:** The corresponding 50%-masked readout accuracy. Thin curves are input-batch means; thick curves are three-batch means.

template, and matched to the aligned edit in Frobenius norm. No edit was trained. The input batch remained the unit of inference, with network initialization and null direction nested within batch.

We evaluated two readout protocols that answer different questions. Under *per-edit refitting*, each edited network received a newly standardized ridge readout fitted on its own clean responses; that readout was then held fixed for masked testing. Under *fixed-readout evaluation*, the unedited low-endpoint standardizer and readout were applied unchanged after the network edit. Table 8 compares the results.

Table 8: **Aligned edit versus equal-norm unaligned edits.** Changes are relative to the unedited low-state endpoint and are percentage points. The final column compares the clean-to-mask loss reduction of the aligned edit with the mean of 16 null edits. Intervals use ten independent input-batch means; three parameter initializations and 16 null directions are nested within batch.

Readout protocol	Δ clean	Δ masked	Loss reduction	Aligned − mean null
Refit on edited clean responses	-0.14	$+10.72$	$+10.86$	$+10.14$
95% CI	$[-0.53, 0.25]$	$[10.03, 11.42]$	$[9.97, 11.76]$	$[9.46, 10.83]$
Original low readout fixed	-15.56	-6.44	$+9.12$	$+8.16$
95% CI	$[-16.31, -14.81]$	$[-6.80, -6.08]$	$[8.38, 9.86]$	$[7.52, 8.81]$

The prespecified loss-gap criterion is positive under both protocols, and its scientific meaning differs between them. With per-edit refitting, the gap shrinks because masked accuracy rises by 10.72 points while clean accuracy is unchanged within uncertainty; the aligned edit produces the largest masked-accuracy gain and loss reduction among the 17 tested directions in every one of the ten batches. With the original readout fixed, both absolute accuracies fall and the gap shrinks only because clean accuracy falls further. The full aligned edit therefore reduces performance with the original readout and supports a more robust readout after clean-only refitting relative to the tested equal-norm unaligned edits.

C.9 Cross-task alignment protocol

The cross-task confirmation used ten held-out input batches and three paired network initializations per batch. For each batch, let P project onto the five leading eigenvectors of the training input-correlation matrix S. Four fixed orientation seeds independently define a unit vector $a_{\|}$ in $\mathrm{range}(P)$ and a unit vector $a_{\perp}$ in its orthogonal complement. At each prespecified alignment level $q \in \{0, .25, .5, .75, 1\}$, the task direction is

$$a(q) = \sqrt{q}\, a_{\|} + \sqrt{1-q}\, a_{\perp}. \tag{20}$$

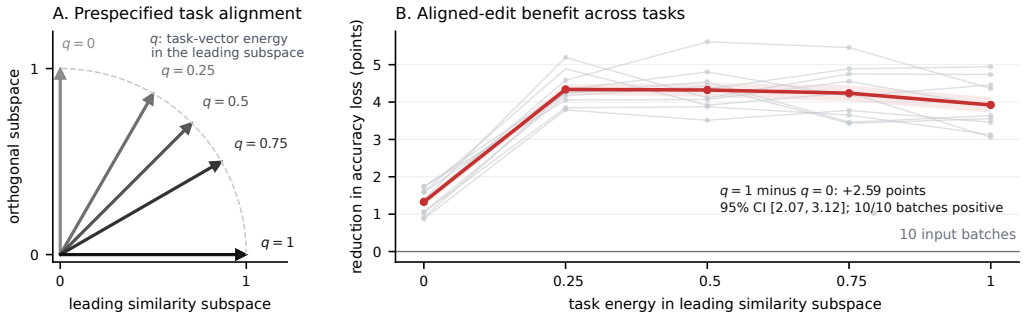


Figure 10: **The functional effect of the endpoint edit depends on task alignment. A:** Binary task directions rotate between the leading five-dimensional input-similarity subspace and its orthogonal complement; q is the squared projection of the unit task vector onto the leading subspace. Task-score variance varies separately. **B:** Reduction in clean-to-masked accuracy loss produced by the aligned recurrent edit relative to the unedited low-state endpoint. Thin curves are means over three nested network initializations and four task orientations within each independent input batch; the thick curve is the mean ± SEM across ten batches. Readouts are fitted separately on complete responses from each representation and held fixed for masked testing.

Clean excitatory inputs are standardized using the 600 fitting examples. The binary label indicates whether their projection onto $a(q)$ exceeds the fitting-split median, giving a balanced fitting target. For the unedited low endpoint, aligned edit, and high endpoint separately, a ridge-linear readout is fitted on complete recurrent responses and held fixed on 600 disjoint complete and 50%-masked examples. Tasks, orientations, and network initializations are nested within input batch.

Here $q = \|Pa(q)\|_2^2$ is the squared projection of the unit task vector onto $\mathrm{range}(P)$. Input variance differs between the leading subspace and its orthogonal complement, so task-score variance and difficulty vary across q.

The primary cell-level statistic is the slope across q of the aligned edit's reduction in clean-to-masked accuracy loss relative to the unedited low endpoint, averaged over the four orientations. The three cell slopes are then averaged within input batch; confidence intervals use the ten independent batch means. The development batch and initialization were excluded before confirmation. All 30 confirmation cells had finite outputs and valid endpoint signs. Table 9 reports the actual curve, which is nonlinear despite the positive prespecified slope.

Table 9: **Cross-task alignment sweep.** Accuracies and paired changes are percentage points. "Low" is the unedited low-state endpoint and "Edit" is the similarity-aligned edit; gap reduction is $(\Delta\text{masked} - \Delta\text{clean})$. Intervals use ten independent input-batch means; three network initializations and four task orientations are nested within batch.

q	Low clean	Edit clean	Δclean [95% CI]	Low masked	Edit masked	Δmasked [95% CI]	Gap red. [95% CI]
0	82.26	82.81	+0.56 [0.32, 0.80]	63.52	65.41	+1.88 [1.71, 2.06]	+1.33 [1.09, 1.57]
0.25	82.79	86.19	+3.41 [3.10, 3.71]	67.04	74.78	+7.74 [7.35, 8.14]	+4.34 [4.02, 4.65]
0.50	82.58	86.33	+3.74 [3.45, 4.03]	67.09	75.16	+8.06 [7.58, 8.54]	+4.32 [3.90, 4.74]
0.75	82.42	86.33	+3.90 [3.54, 4.27]	67.07	75.21	+8.14 [7.61, 8.67]	+4.24 [3.75, 4.72]
1	82.19	86.30	+4.11 [3.62, 4.61]	67.19	75.22	+8.03 [7.55, 8.52]	+3.92 [3.45, 4.40]

The prespecified batch-mean slope was positive in 10/10 batches (mean 0.0204 accuracy-loss units per unit q, 95% CI $[0.0151, 0.0256]$). The edit increased both clean and masked accuracy at every sampled q, with the larger masked gain producing a positive gap reduction in all ten batches. This supports task dependence within the tested family. The plateau from $q = 0.25$ onward, together with unmatched task-score variance, precludes interpreting the fitted slope as a mechanistic linear law.

C.10 Fixed-window history check

A separate experiment asks how quickly a learned difference decays after the learning condition changes. We trained paired networks toward the same final low or high learning condition, starting

either from that condition or from the opposite one. All branches received a fresh AdamW state. Within each final condition, the two histories shared the same 600-update noise stream and duration. At several checkpoints, we froze every network and evaluated it with the same $g_{\text{probe}} = 0.5$ and held-out inputs. Main Fig. 3A–C reports the protocol and structural trajectories; Fig. 11 reports the accompanying nonstructural readouts. The analysis covers this specified adaptation window and leaves convergence to distinct solutions open.

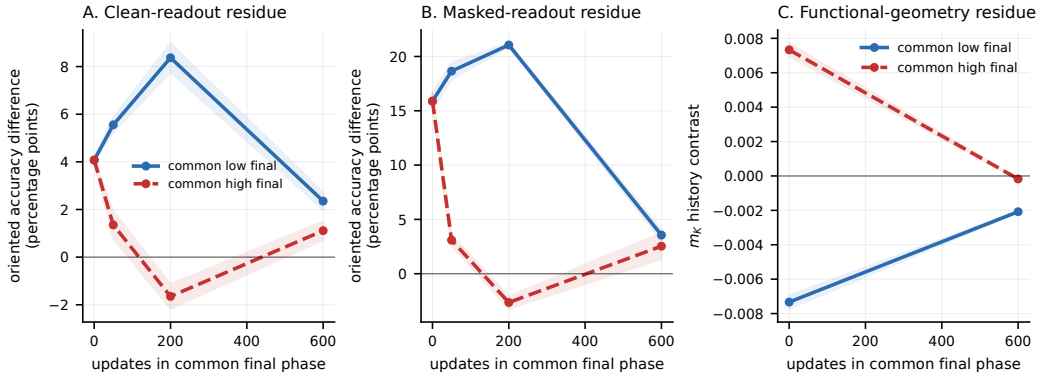


Figure 11: **Nonstructural readouts during matched final adaptation. A,B:** Oriented differences in clean- and 50%-masked readout accuracy for the four cohorts whose protocol and structural trajectories appear in Fig. 3A–C. Positive values indicate retention of the source-condition bias. **C:** The functional-slope history contrast follows the low-final residue but has the opposite sign to the high-final structural and readout directions. Curves show means ± SEM over three independent input batches.

The LL, HL, HH, and LH branches start from paired low- and high-trained sources. All branches use newly initialized AdamW objects; within a common final learning condition, the switched and constant branches receive the same final-phase training-noise sequence. Frozen evaluations occur at updates 0, 50, 200, and 600 using the common probe and paired held-out examples. In Table 10, positive accuracy values denote retention of the source-condition bias: HL−LL for a common low final condition and HH−LH for a common high final condition.

Table 10: **History residue after 600 updates under matched final learning conditions.** Structural rows summarize the ten-batch extension; nonstructural rows use the original three-batch confirmation because these readouts were outside the frozen extension. Paired parameter initializations are nested within batch. Accuracy contrasts are percentage points.

Common final	Oriented contrast	Mean	95% CI	Batch signs
Low	Structural m_W: HL−LL	-0.0116	$[-0.0141, -0.0091]$	10/10 negative
	Clean accuracy: HL−LL	$+2.35$	$[0.78, 3.92]$	3/3 positive
	Masked accuracy: HL−LL	$+3.56$	$[1.39, 5.72]$	3/3 positive
High	Structural m_W: LH−HH	$+0.280$	$[0.268, 0.291]$	10/10 positive
	Clean accuracy: HH−LH	$+1.11$	$[-0.57, 2.79]$	3/3 positive
	Masked accuracy: HH−LH	$+2.54$	$[-2.69, 7.76]$	3/3 positive

Both high-final accuracy directions are uncertain, and the high-final functional-slope contrast is opposite to the structural and accuracy orientation. The common-low masked-accuracy difference also decays from 21.06 points at update 200 to 3.56 at update 600 while structural slopes continue to move. The functional readouts therefore lack a shared, sign-preserving correspondence with the structural history contrast across the two final conditions. The direct m_W measurement establishes the structural residue, and the longer audit below tracks its evolution while leaving convergence open.

C.11 RECIPROCAL EXCHANGE OF THE RETAINED HISTORY COORDINATE

We prespecified a reciprocal intervention for the common-low LL/HL pair, the one final condition in which the structural and masked-readout history contrasts were both reproducible. The analysis used ten independent correlated-input batches (seeds 1302–1311) with three paired parameter initializations nested within each batch. Update 600 was primary and update 200 secondary. All 30 primary common-low cells passed the finite-output and nonzero-response checks. The stored history checkpoints reproduced the previously frozen structural trajectories exactly before any edit was applied.

The decomposition and reciprocal edits are defined in Eq. 3. The achieved target slope differed from its paired target by at most 1.9×10^{-8}. We evaluated the four history–coordinate combinations Y_{00} (LL), Y_{01} (LL implant), Y_{10} (HL erase), and Y_{11} (HL). The symmetric coordinate effect and complementary background effect are

$$C = \tfrac{1}{2}[(Y_{01} - Y_{00}) + (Y_{11} - Y_{10})], \qquad B = \tfrac{1}{2}[(Y_{10} - Y_{00}) + (Y_{11} - Y_{01})], \tag{21}$$

so the total history contrast satisfies $Y_{11} - Y_{00} = C + B$ exactly. This four-cell decomposition quantifies coordinate and background contributions under the two edited backgrounds. Natural indirect effects and complete causal mediation require additional assumptions and interventions. The symmetric decomposition averages each factor's effect across the two backgrounds. In particular, B differs from either single-direction residual and aggregates the unexplained contribution without identifying its remaining parameter sources.

The primary readout protocol standardizes each edited representation and fits a ridge-linear decoder to its complete responses, then holds that decoder fixed on paired 50%-masked inputs. A secondary protocol carries the original branch-specific standardizer and decoder through the edit. Sixteen null exchanges were constructed from eight prespecified random symmetric directions and both signs, projected orthogonal to $\widetilde{S}$ and scaled to the same Frobenius norm as the aligned exchange. Their largest relative norm mismatch was 2.9×10^{-9}. Confidence intervals and sign counts use ten input-batch means; parameter initialization and null direction remain nested within batch.

Table 11: **Common-low history-component exchange at update 600.** Accuracy effects are percentage points under per-edit clean-only readout refitting. The implant and erase rows are oriented so a positive value closes the LL–HL masked-performance gap. Intervals use ten independent input-batch means.

Contrast	Mean	95% CI	Batch signs
Total HL−LL, masked	$+3.74$	$[+2.83, +4.66]$	10/10 positive
LL implant, masked	$+1.18$	$[+0.72, +1.64]$	10/10 positive
HL erase, masked loss	$+0.72$	$[+0.42, +1.01]$	10/10 positive
Symmetric coordinate C, masked	$+0.95$	$[+0.61, +1.29]$	10/10 positive
Symmetric coordinate C, complete	$+0.06$	$[-0.07, +0.18]$	5/10 positive
Masking-specific C	$+0.89$	$[+0.62, +1.16]$	10/10 positive
Aligned $C-$ mean null C, masked	$+0.97$	$[+0.68, +1.26]$	10/10 positive
Complementary background term B, masked	$+2.80$	$[+2.07, +3.53]$	10/10 positive

Table 12: **Absolute four-cell accuracies at update 600.** Values are percentages with 95% confidence intervals over ten independent input-batch means; three paired parameter initializations are nested within each batch. "Refit" fits a clean-only readout separately to each representation. "Fixed" carries the original branch-specific standardizer and decoder through the edit.

Condition	Refit complete	Refit masked	Fixed complete	Fixed masked
Y_{00}: LL	76.97 [76.04, 77.91]	57.07 [56.07, 58.08]	76.97 [76.04, 77.91]	57.07 [56.07, 58.08]
Y_{01}: LL implant	77.10 [76.24, 77.96]	58.25 [57.06, 59.45]	76.04 [75.19, 76.90]	56.65 [55.61, 57.69]
Y_{10}: HL erase	79.59 [78.72, 80.47]	60.10 [58.68, 61.52]	80.58 [79.66, 81.51]	61.44 [59.97, 62.92]
Y_{11}: HL	79.58 [78.64, 80.51]	60.82 [59.36, 62.27]	79.58 [78.64, 80.51]	60.82 [59.36, 62.27]

The aligned exchange ranked first among the 17 directions in every batch. At update 200, the coordinate effect was larger ($+3.68$ points, 95% CI $[3.10, 4.27]$), consistent with the larger total history contrast earlier in adaptation. With branch-specific decoders fixed at update 600, implanting the HL coefficient into LL reduced masked accuracy by 0.42 points ($[0.25, 0.59]$) and complete-input

accuracy by 0.93 points ($[0.71, 1.14]$). Conversely, replacing the coefficient in HL with the LL value increased masked accuracy by 0.63 points ($[0.45, 0.81]$) and complete-input accuracy by 1.01 points ($[0.62, 1.39]$). Their oriented average was therefore -0.53 points for masked inputs and -0.97 points for complete inputs. Positive reciprocal closure therefore requires clean-only readout refitting and reverses with fixed deployed decoders.

The common-high exchange serves as a scope check because the original high-final nonstructural contrasts lacked alignment with m_W. In input batch 1308, initialization 2304, all four update-600 outputs were finite, but their nonzero-response fractions were 97.83–97.88%, below the prespecified 98% validity threshold. The entire batch was therefore excluded from the complete paired ten-batch summary. Among the nine remaining batches, the refitted masked coordinate effect was positive in four and negative in five, while the masking-specific effect was opposite to the common-low result in all nine. The evidence therefore supports a partial, background-dependent contribution in the common-low pair. Most of the total history-dependent performance difference resides in other parameters, the orthogonal recurrent background, or their interaction with the refitted readout.

C.12 Longer matched-objective and stationarity audit

We reran the four history branches for ten independent correlated-input batches and three paired initializations per batch. All branches used fresh AdamW states, paired noise within each final condition, a common maximum of 4,800 final-phase updates, and a fixed four-noise bank for evaluating objective components and gradients. Stationarity at update 4,800 required all three prespecified conditions: in both the 2,400–3,600 and 3,600–4,800 intervals, the absolute fixed-bank objective change had to be at most 2% of the branch's absolute 0–4,800 objective change; normalized parameter drift had to be at most 0.5% in each interval; and the fixed-bank gradient norm at update 4,800 had to be at most 10% of its update-0 value.

The drift is computed in the space of raw trainable tensors (before the softplus transformations used by constrained blocks):

$$d_\theta(t_1, t_2) = \frac{\|\theta(t_2) - \theta(t_1)\|_2}{\frac{1}{2}[\|\theta(t_1)\|_2 + \|\theta(t_2)\|_2]}. \tag{22}$$

The gradient diagnostic is the Euclidean norm, over all raw trainable tensors, of the gradient of the mean total objective across the four fixed noise realizations. It is measured before clipping and compared with the same branch's own update-0 value. These operational thresholds diagnose slowing over the sampled trajectory; they provide no mathematical test of an exactly zero gradient.

The separate *matched-objective* comparison searched all cross-history checkpoint pairs at updates ≥ 600, selected the pair with the smallest absolute difference in fixed-bank total objective, and accepted it only when the mismatch was at most 1% of the pooled objective range across both trajectories. The pooled range uses all recorded checkpoints from update 0 through 4,800, and the matching criterion excludes the structural slope. All 60 cell-by-final-condition matches were accepted. Matching one scalar objective leaves loss components, gradients, parameters, and optimization progress free to differ. It complements the fixed-time contrast by comparing similar values of one scalar objective.

Table 13: **Actual matched-objective values.** Ranges cover the 30 cell-level matches per final condition. The complete pairs $(t_c, t_s, J_c, J_s, |J_c - J_s|, m_{W,c}, m_{W,s})$ are retained in the released result file.

Final	t_c	t_s	J_c range	J_s range	Max. $\lvert J_c - J_s \rvert$
Low	600–3,600	2,400–4,800	0.08192–0.09592	0.08193–0.09591	6.25×10^{-6}
High	600	4,800	0.18925–0.22982	0.18936–0.22986	1.86×10^{-4}

The maximum mismatch divided by the corresponding full-trajectory pooled range was 4.45×10^{-5} for low-final and 3.05×10^{-5} for high-final matches.

All 30 cells were valid, and both same-update contrasts had the displayed direction in every batch at updates 600 and 4,800. Their later behavior was asymmetric: the low-final difference nearly decayed, whereas the high-final structural projection expanded as the constant-high branch continued to move (Fig. 12). Of the 120 branches, 90 passed the objective-change test, 69 passed the gradient test, and

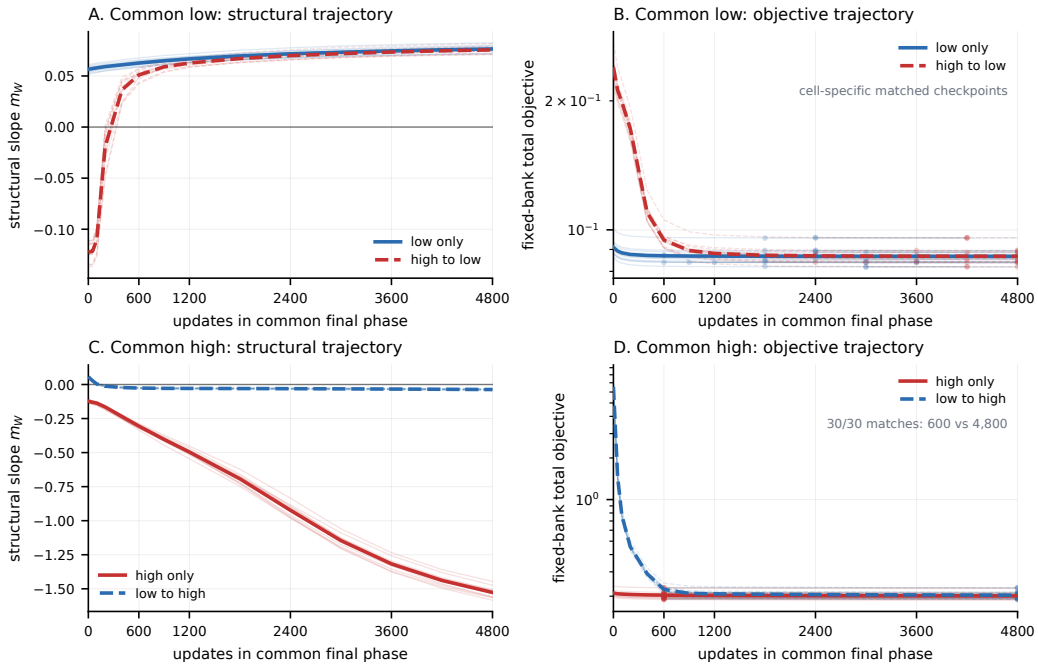


Figure 12: **Longer history trajectories and matched-objective checkpoints. A,C:** Structural slope m_W during common low and high final learning. **B,D:** Fixed-bank total objective on a logarithmic scale. Thin curves are independent input-batch means, thick curves are grand means, and shaded bands are ±SEM over ten batches; three paired initializations are nested within each batch. Faint horizontal segments join the two cell-specific checkpoints selected by objective matching. Low-final checkpoint pairs vary across cells; all 30 high-final matches compare the constant-history branch at update 600 with the switched-history branch at update 4,800.

Table 14: **Longer structural-history audit.** Contrasts are switched minus constant history. Intervals use ten input-batch means; three initializations are nested within batch. The matched-objective comparison may use different update counts.

Common final	Comparison	Mean	95% CI	Batch signs
Low: HL−LL	Update 600	−0.01162	[−0.01412, −0.00911]	10/10 negative
	Update 4,800	−0.00094	[−0.00132, −0.00056]	10/10 negative
	Matched objective	+0.00402	[+0.00298, +0.00506]	10/10 positive
High: LH−HH	Update 600	+0.27968	[+0.26847, +0.29088]	10/10 positive
	Update 4,800	+1.48985	[+1.45785, +1.52184]	10/10 positive
	Matched objective	+0.26889	[+0.25855, +0.27924]	10/10 positive

none passed the parameter-drift test; consequently no branch met all three stationarity criteria. The matched-objective comparison preserved the high-final direction but reversed the low-final contrast. These comparisons show that the measured projection depends on both exposure history and the point along the subsequent optimization trajectory. The saved audit records m_W and total parameter norms without separately recording off-diagonal W_{EE} variance and correlation with S, so growth of m_W has an ambiguous decomposition. The tested horizon leaves eventual convergence open.

Warm starts can alter generalization even when neural networks reach similar training losses (Ash & Adams, 2020), making finite-time path dependence a broader property of learning systems. Consolidation and continual-learning models add persistent variables or protected parameters (Fusi et al., 2005; Benna & Fusi, 2016; Kirkpatrick et al., 2017). The present finite-window result is compatible with both ordinary warm-start dynamics and these broader adaptive-memory mechanisms.

The explicitly state-gated Oja-style comparator in Appendix G provides a second existence example of the qualitative switching effect. Its scope is limited to the implemented comparator.

D ADDITIONAL ENDPOINT AND READOUT CHECKS

An early development pair used prototype cosine similarity for completion and pair-separation probes. Pairwise cosine separation had the opposite predicted sign, so these non-diagnostic metrics were replaced before any confirmation seed was evaluated. A later baseline audit classified raw missing-coordinate gain as non-diagnostic for computational benefit. Both development results are excluded from the confirmation analyses.

This endpoint-replication cohort is separate from both the original three-batch learning-state confirmation and the later ten-batch structural-history and aligned-edit extensions. It uses three independent input batches with five paired initializations per batch. For each pair, the evaluation generator draws 1,200 new samples and records the index of the largest of five Gaussian latent factors as the label. A deterministic paired Bernoulli mask hides half of the excitatory coordinates on average. The first 600 clean samples fit a ridge-linear readout with intercept after standardization on that split; the remaining 600 masked samples are used once. Endpoint validity requires finite output and nonzero responses for at least 98% of clean and masked samples. All 15 endpoint pairs were valid.

Table 15: **Common-probe endpoint computation.** Means are over three independently generated input-batch means; five paired network initializations are nested within each batch. Confidence intervals use a t interval across batches. Accuracy differences are percentage points.

Quantity	Mean	95% CI	Pair sign
High minus low masked accuracy	16.11	$[10.01, 22.21]$	15/15
High minus input masked accuracy	13.58	$[7.60, 19.55]$	15/15
Low minus high structural slope m_W	0.18121	$[0.16999, 0.19244]$	15/15

Every endpoint was valid, all 15 paired masked-accuracy and structural contrasts had the displayed signs, and their intervals across independent batch means excluded zero. The originally prespecified raw-gain contrast also met its sign criterion; Appendix C.3 shows that output-scale dependence prevents a useful-reconstruction interpretation. Across paired differences, the structural contrast correlated descriptively with the masked-accuracy contrast ($r = 0.71$). Mediation requires the intervention reported separately.

E SWITCHING ROBUSTNESS AND SCOPE

The reference assay trains paired low- and high-state endpoints for 600 updates from the same initialization and paired denoising-noise stream. Each endpoint then continues for 1,200 updates with its complete AdamW state retained. The opposite-direction gates follow the same immediate 20-update cosine ramp: onset at update 0, midpoint at update 10, and target gate at update 20, with no pre-ramp dwell. The two directions share the switch-phase noise stream. Zero-crossing time is counted in actual updates from ramp onset and is linearly interpolated between structural-slope samples recorded every ten updates.

Table 16: **Schedule-matched $N = 100$ switching.** Times are updates from ramp onset. Initializations are averaged within independently generated input batches; values are mean $\pm$ SEM over three batches.

Input family	$\tau_{\text{coop}\to\text{comp}}$	$\tau_{\text{comp}\to\text{coop}}$	Predicted order
Correlated	24.03 ± 1.47	329.98 ± 27.37	9/9
Gabor	23.50 ± 0.03	665.54 ± 74.06	9/9
Natural images	15.62 ± 0.55	785.55 ± 45.56	9/9

An alternative Gabor tiling and natural-image patches from the disjoint scikit-image examples `chelsea`, `moon`, and `brick` each retained the order in 9/9 additional matched runs. At fixed

$N = 100$, equal E/I counts, three learning rates, removal of gate-dependent sparsity, and removal of sparsity altogether each retained the order in all five tested paired initializations per condition, while changing the magnitude. Equal counts were tested as 50E/50I at fixed total N with both the original per-synapse inhibitory scale and aggregate-inhibition matching; the corresponding ratios were 6.78 and 4.46, compared with 11.02 in the 80E/20I reference. Because population composition and inhibitory normalization change together, the causal role of cell count alone remains unresolved.

Zero-crossing time depends on the starting and destination values of m_W as well as on the learning dynamics. Even two exponential relaxations with the same intrinsic time constant can cross zero at different times when their endpoint magnitudes differ. The observed order therefore describes this protocol's structural crossing sequence. Identifying intrinsic adaptation rates requires a separate dynamical model.

Table 17: **Optimizer-state reconciliation.** Each ratio belongs to a distinct protocol. The retained/reset pair directly isolates stored AdamW state.

Analysis	Ratio	Relevant difference
Independent batches	13.98 ± 2.05	3 batches $\times$ 3 initializations
Paired retained state	10.99 ± 0.05	one batch, 5 paired seeds
Paired reset state	1.408 ± 0.027	same endpoints; AdamW state cleared
Initial illustration	7.80 ± 0.12	unequal prehistory/window; superseded

Adam, AdamW, and momentum SGD reached the endpoint sign split after optimizer-specific calibration. Plain SGD left the cooperative low-gate endpoint unresolved within the tested 1,200-update budget, so optimizer-independent endpoint behavior remains open. The network-size control used equal 4,800-update windows and the prespecified scaling $\mathrm{lr}(N) = \mathrm{lr}(100)\sqrt{N/100}$. All three runs at $N = 500$ and $N = 800$ retained the $N = 100$ order. At $N = 1000$, competition-to-cooperation crossed after 349 ± 10 updates while cooperation-to-competition remained right-censored. This result establishes a failure of generalization under one calibration; the simultaneous numerical changes with N prevent a causal size interpretation.

The originally illustrated switching branches used total lengths of 1,200 and 4,800 updates. Both gates changed over 20 updates and time was counted in actual updates, but the branches differed in pre-switch training, accumulated AdamW state, and observation window. We retain that unmatched result only as development history and use the schedule-matched experiment above for all switching-time claims.

F Reduced gradient identity and the activity second moment

The following identity concerns a stable linear equilibrium reduction. Gradients of the implemented network are computed separately through its finite-step unroll. Let $R = (I - W_{\mathrm{eff}})^{-1}$ and $\mathbf{r}_E^* = R\mathbf{u}_{\mathrm{eff}}$. Matrix differentiation gives

$$\frac{\partial r_{E,k}^*}{\partial W_{EE,ij}} = R_{ki} r_{E,j}^*, \qquad \frac{d\mathcal{J}}{dW_{EE,ij}} = \left[R^\top \nabla_{\mathbf{r}_E^*} \mathcal{J}\right]_i r_{E,j}^*. \tag{23}$$

Thus cross-neuron activity products enter the instantaneous gradient. We summarize them with $M = \mathbb{E}[\mathbf{r}_E^* \mathbf{r}_E^{*\top}]$ and $M_{\mathrm{off}} = M - \mathrm{diag}(\mathrm{diag}\, M)$. We use M_{off} as a diagnostic. The paired optimizer-reset comparison starts from identical weights, activity, and M_{off} yet produces different transition times, ruling out $\|M_{\mathrm{off}}\|_F$ as a sufficient recovery-time clock.

G State-gated Oja-style comparator

The comparator learns only the recurrent E-to-E block, fixes fast inhibitory gain at $G = 0.2$, and uses a direct update without Adam or momentum:

$$\Delta W_{ij} = \eta\big[(1 - 2g)C_{ij} - \kappa C_{ii} W_{ij}\big], \tag{24}$$

where C is centered activity covariance. Three exploratory runs selected $\eta = 0.02$ and $\kappa = 8$ and were excluded. With the rule fixed, three new correlated-input batches and three initializations per

batch all reached both endpoint signs and crossed zero in both directions; batch-level crossing times were 66.55 ± 2.53 and 76.90 ± 1.52 updates. Structural history contrasts under common final learning conditions had the predicted signs in all nine runs during the prespecified 10–200-update window and largely decayed by update 600.

The comparator is an explicitly state-gated recurrent Oja-style variant. Its gate controls the sign of plasticity, and normalization may contribute to the direction. It provides a second existence example of the qualitative effect beyond SG-EAP, without characterizing canonical Oja, BCM, or adaptive rules as a class.

H BIOLOGICAL OBSERVABLE AND PROSPECTIVE TEST

The mouse V1 dataset in Russell et al. (2024) tests only whether an interaction–correlation slope changes along a current contrast axis. An independent grating block defines tuning similarity. Photostimulation measures influence on non-target cells more than $25\,\mu$m from the nearest target. Within each of nine sessions from nine mice, influence is regressed on nearest-target tuning similarity at each of six contrasts; session is the replicate. As a hierarchical check, the observer-level model is

$$\text{influence} \sim S + \text{contrast} + S{:}\text{contrast} + (1 + \text{contrast} \mid \text{session}), \tag{25}$$

where numeric contrast uses the fractions $\{0, .01, .02, .05, .10, 1\}$, and session/mouse is the random-effect grouping variable with a random intercept and contrast slope. The REML fit uses 142,296 observer-level rows in nine groups. The reported statistic is the coefficient divided by its model-based standard error; the two-sided p-value uses the large-sample Wald normal approximation. The mean session-level slope changes from $+0.009$ at 0% contrast to -0.013 at 100%. The specified mixed model gives a negative similarity-by-contrast interaction ($\beta = -0.015$, Wald $z = -3.16$, normal-reference $p = 1.6 \times 10^{-3}$; Fig. 13). The nine session/mouse groups and the random-effects structure support treating individual session trajectories and effect sizes as the primary descriptive evidence; the model-based p-value provides a covariance-structure-dependent hierarchical check. Fixed-weight gain models can also produce this sign change, so the dataset establishes current-drive consistency while leaving learned history untested.

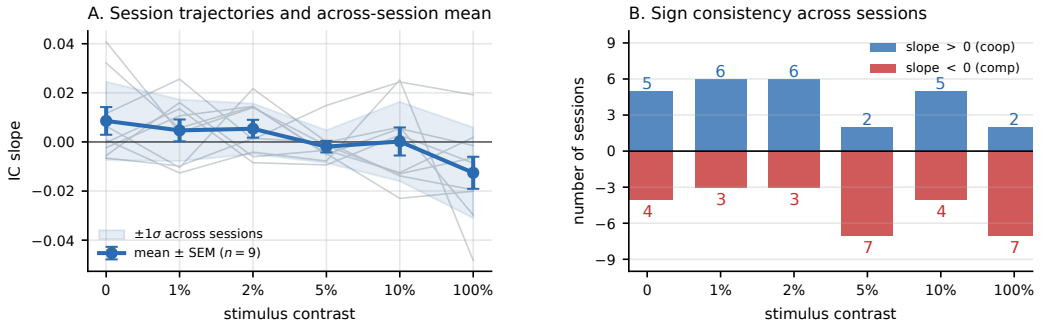


Figure 13: **V1 interaction slopes change with current contrast. A:** Thin lines are the nine session-level trajectories (nine mice), the shaded band is ± 1 SD across sessions, and blue markers show the mean $\pm$ SEM at each of six contrasts. **B:** Numbers of positive and negative session slopes. The available data sample current contrast across sessions; matched histories and longitudinal recovery require a prospective experiment.

A prospective biological experiment could use longitudinal perturbation maps after matched low-only versus high-to-low exposure and high-only versus low-to-high exposure, with present stimulus, behavioral state, time since transition, and perturbation protocol held fixed. Section 4 establishes bidirectional finite-window history for the structural slope m_W; the common-high history contrast in the functional slope m_K has the opposite sign. The primary biological question is therefore whether matched-history groups differ at all in their causal interaction maps. A lower interaction–similarity slope after high-to-low than low-only exposure and a higher slope after low-to-high than high-only exposure would support an additional structure-to-function sign generalization. A prespecified equivalence result would weigh against a functional history effect in that window, whereas opposite

directions would weigh against the sign generalization. Either outcome would leave the structural m_W result unchanged.