THE SPECTRAL SHORTCUT: HOW A FAST CONTEXTUAL PATHWAY SUPPRESSES SPECTRAL LEARNING IN SERIAL SPECTRAL–SPATIAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper we study shortcut learning in serial spectral–spatial models for infrared tissue classification: an encoder processes each pixel's spectrum, and a contextual head reads the encoded neighbourhood. Changing their relative training rates can alter the learned representation while leaving the fitted classifiers similarly context-dependent. In a controlled synthetic experiment, slowing the head by a factor of 256 raises mean linear-probe accuracy from 0.65 to 0.83 at matched training loss, while original-classifier accuracy under contextual reversal changes only from 0.12 to 0.14. Retraining a fresh head for $20{,}000$ steps on decorrelated-context data, with the encoder frozen, yields a mean paired reversal-accuracy gain of 0.17 for slow-head over fast-head encoders, positive in all nine evaluable pairs; one of ten prespecified slow-head runs does not reach the matching loss. The mean recovery advantage persists at all tested combinations of three learning rates and three budgets up to $80{,}000$ steps. A solvable logistic model establishes a sufficient mechanism: contextual readout width and learning rate jointly control a bound on spectral adaptation, with reversal failure under explicit initialization and finite-width conditions. Readout normalization removes the width dependence without necessarily preventing failure. A two-layer ReLU encoder reproduces the accessibility and recovery orderings. Constructed-context experiments with infrared tissue spectra and public hyperspectral spectra reproduce suppression and reliance, but the effects of slowing the head vary across settings. In the high-amplitude tissue construction, blocking neighbour-derived encoder gradients restores strong accessibility and recovery responses to slowing the head. In the tested natural 3×3 neighbourhoods, reliance occurs without a detected encoder-accessibility deficit. Together, these results distinguish contextual reliance from the usefulness of the learned representation for subsequent head retraining.

1 INTRODUCTION

Machine learning offers automated analysis of stained tissue images, and infrared imaging is being pursued to add label-free molecular information alongside spatial structure (Berisha et al., 2019): infrared microscopy, including quantum-cascade-laser (QCL) instruments that image tissue microarrays within clinical timeframes (Ferguson et al., 2025), measures a spectrum at each pixel. Combining them is not straightforward, and how much spectral information a spatial–spectral classifier learns to use is unclear. Müller et al. (2023) report strong spectral compression with little performance loss, and pixel-level deterioration under spatial permutation while region-level classification stays accurate; O'Leary et al. (2026) find that a learned sixteen-feature bottleneck preserves performance across architectures, including a nonspatial MLP. These observations establish neither that spectral information is redundant nor that a spatial classifier has learned a useful centre-spectrum representation: compact spectral representations, contextual reliance, and suppressed spectral learning remain distinct possibilities, and fitted accuracy does not tell them apart.

We study whether rapid contextual fitting can suppress the learning of a useful centre-spectrum representation, the molecular signal that motivates chemical imaging. Our models are serial: an encoder maps each pixel's spectrum to features, and a spatial head reads the encoded neighbourhood (Berisha et al., 2019; O'Leary et al., 2026). We call evidence in the target pixel's own spectrum, the

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.

centre of the neighbourhood, the *spectral cue* and evidence in its neighbours' spectra the *contextual cue*. This is a shortcut-learning question about competing cues and their availability (Hermann et al., 2024), with relative module rates as the intervention. A solvable model gives one sufficient mechanism (Section 3). Controlled experiments then separate what the encoder learns, what the fitted classifier uses and what a fresh head can recover (Sections 4 and 5). In the synthetic model, similarly context-reliant classifiers yield different representation recoverability at matched training loss. The ordering reverses at equal joint-training updates, making the comparison point part of the finding (the $80{,}000$-step sensitivity varies only the retraining budget, encoders fixed). In the tested natural neighbourhoods, reliance occurs without a detected accessibility deficit.

Contributions.

- **An exact sufficient mechanism.** In a serial logistic model with M replicated readouts, gradient flow moves the spectral coefficient at the matched margin by at most a fraction $1/(1 + Mv_0^2)$ of the update a spectral-only model needs, and the fitted model fails under reversal under explicit initialization and finite-width conditions; a $M^{-1/2}$-normalized readout removes the width dependence without necessarily removing the failure (Theorem 1).
- **Recoverability depends on the comparison point.** In a synthetic model, slowing the head by $256\times$ raises the spectral cue's linear accessibility from 0.65 to 0.83 at matched loss while the classifiers' reversal accuracy moves only from 0.12 to 0.14; at that matched fit a fresh head retrained on decorrelated context gains a mean paired $+0.17$ in reversal accuracy for slow- over fast-head encoders (nine evaluable pairs), but after $15{,}000$ or $40{,}000$ equal joint updates it recovers more from the fast-head encoder (-0.12 and -0.14, ten pairs each; Figure 2); a two-layer ReLU encoder reproduces the matched-loss orderings (Section 4).
- **Scope across measured spectra.** With infrared tissue spectra and three public hyperspectral datasets in constructed neighbourhoods, informative context is relied on in every regime and suppresses spectral adaptation where the cue is not already accessible. Rate responses vary across the measured-spectrum settings. Blocking neighbour-derived encoder gradients restores accessibility ($0.63 \to 0.91$) and recovery ($0.53 \to 0.91$) at matched fit in the high-amplitude tissue construction but not on HeiPorSPECTRAL. In natural 3×3 neighbourhoods reliance occurs without a detected accessibility deficit (Section 5; Appendices E, F).

Related work. One predictive feature suppressing another is gradient starvation (Pezeshki et al., 2021), a shift-fragile bias toward simple features is simplicity bias (Shah et al., 2020), and modality competition (Wang et al., 2020; Peng et al., 2022; Huang et al., 2022; Du et al., 2023). Huang et al. (2022) prove modality competition and incomplete feature learning in a late-fusion network, and Izmailov et al. (2022) show how training and design choices change core-feature decodability under spurious reliance; our increment is a controlled within-architecture module-rate intervention whose recoverability ranking depends on the comparison point (matched loss versus equal updates), with an observed contrast between constructed and tested natural contexts. Core features can survive unused while a spurious one is used, and last-layer retraining restores robustness (Kirichenko et al., 2023); we ask how the original relative module rate changes that recovery, retraining the whole head from a frozen encoder against a matched-fit control. Ye et al. (2023) give conditions under which core features are learned too poorly for later probing: that distinction is not ours alone. Kumar et al. (2022) show a random head distorting *pretrained* features at fine-tuning time; ours is from-scratch suppression with a fixed direct spectral readout. Barsbey et al. (2025) vary a *global* rate through bias-conflicting examples; ours is a relative module rate. A common rate only rescales time under ideal gradient flow; Mirzaie et al. (2026) show that finite-step training adds an $O(\epsilon)$ mini-batch-variance regularizer reducing spurious reliance, whereas our factor $1/(1 + \kappa M v_0^2)$ (κ the contextual rate; Section 3) survives the small-step limit in which that term vanishes. Faghih et al. (2026) correct spatially disjoint evaluation and receptive fields; diagnosing which cue the fitted representation permits is complementary. Neither the compression nor the permutation result (Müller et al., 2023; O'Leary et al., 2026) measures what the encoder holds. At one bottleneck dimension our rate interventions give similarly context-reliant predictors whose frozen encoders support different shifted accuracy, and a learned twelve-dimensional bottleneck matches the model without a bottleneck while sixteen principal components do not (Appendix E). We use the gradient-flow balance invariant in linear networks (Du et al., 2018; Saxe et al., 2014; Yun et al., 2021) and lazy

training (Chizat et al., 2019; Yang & Hu, 2021). Xu & Ziyin (2025) solve a one-hidden-layer linear network with layer-wise rates exactly from the same balance (squared loss, one data direction, no directly read coordinate); that fixes the trajectory, not the quantity we bound, the displacement of a fixed-readout coefficient at a prescribed cross-entropy margin: our contribution is the consequence, reliance and its recoverability, not the dynamics. normalization-sensitive (Karakida et al., 2019; van Laarhoven, 2017) (Appendix G); freezing in practice (Coil & Cheney, 2025) is the setting, not a remedy.

2 Model, task, and what counts as failure

Serial spectral–spatial model. An input is an image whose pixel p carries a spectrum $x_p \in \mathbb{R}^S$ and a label $y_p \in \{-1, +1\}$. A spectral encoder $f_{\boldsymbol{\theta}} : \mathbb{R}^S \to \mathbb{R}^K$ acts on each pixel separately, and a spatial model $g_{\boldsymbol{\phi}}$ of width M reads the encoded neighbourhood of p to produce a logit F_p (Figure 1). The encoder is linear, $f_{\boldsymbol{\theta}}(x) = Wx$ with $W \in \mathbb{R}^{K\times S}$, in the theorem and in the primary experiments, with a two-layer ReLU encoder extension in Section 4; the spatial model is a replicated readout in the theorem, a ReLU convolutional head in the synthetic experiment and a ReLU patch head in the real-spectrum experiment. Training minimizes the mean margin loss $\mathcal{L} = \frac{1}{N}\sum_p \log(1 + e^{-y_p F_p})$ by gradient flow or plain gradient descent with a base rate η; the baseline uses that one rate for every parameter, and the interventions multiply the rate of a named block by a stated factor.

Two cues. The training distribution carries the label in two places. The *spectral cue* lies in the pixel's own spectrum along a direction the encoder must discover: in the constructed low-accessibility settings a random encoder attenuates it. The *contextual cue* lies in the neighbourhood along directions of large amplitude: a random encoder passes it, so a trained readout can use it at once. Accessibility in this sense is the availability of Hermann et al. (2024), measured here as the linear readability of a cue from the encoder's output at initialization and after training. The two cues are calibrated to matched discriminant accuracy (synthetically by signal-to-noise; on real spectra by a training-side estimate), so preferring one is not preferring the more predictive one.

Matched fit, spectral learning, reliance, failure. We compare models at their first attainment of a prespecified training-fit threshold: a common margin in the solvable model and mean training loss in the experiments. Spectral learning means adaptation of the designated spectral coefficient in the theorem and increased linear accessibility of the centre's spectral cue in the experiments. Contextual reliance is the change in predictions or risk when the assigned contextual signal changes while the target's spectral cue, label and noise are preserved; it is a failure when the shifted risk exceeds that of a same-architecture comparator trained with decorrelated context to the same threshold. The theorem supplies an exact reversal failure against a spectral-only model; the experiments measure the risk difference against the trained comparator, assuming no zero-error reference.

Test conditions. Every evaluation family is generated from the same labels and noise as *iid* (context agrees with the label), *reversed* (the assigned context carries the opposite label) and *context-random* (an independent label: context present but uninformative).

3 Main result: selective spectral learning at matched fit

We state the mechanism in the simplest serial model that has both cues, a shared encoder, and a replicated contextual readout. Training inputs are two sites: the pixel's own spectral coordinate S and a contextual coordinate C, both equal to the label. The encoder $W = (a, v)$ reads each site with one coefficient; the head adds the spectral coordinate directly and the contextual one through M replicated readouts. Replication is the only way width enters. Theorem 1 is proved in Appendix A; the closed forms are verified numerically in Appendix B (Appendix Figure 4).

Theorem 1 (Selective spectral learning at matched fit)**.** *Let Y be uniform on $\{-1, 1\}$, $x_1 = (S, 0)$, $x_2 = (0, C)$, and $S = C = Y$ in training. The shared encoder $W = (a, v)$ produces $z_1 = aS, z_2 = vC$; the head is $F = z_1 + bz_2$, $b = \sum_{j=1}^{M} \beta_j$. For integer $M \geq 1$, train (W, β) by unit Euclidean*

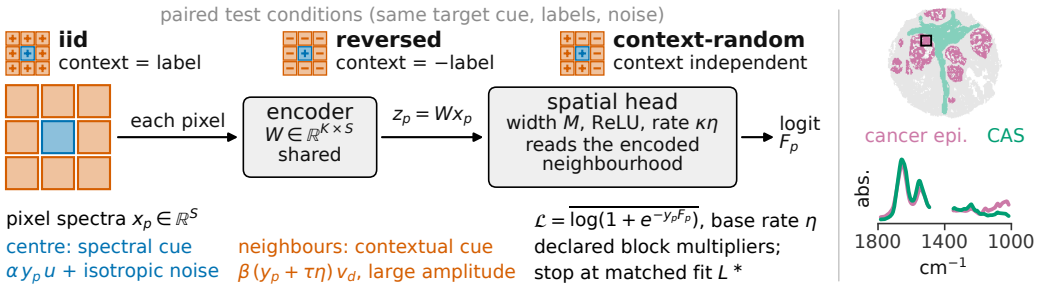

Figure 1: Left: the serial model and the two cues. In the controlled synthetic experiment at matched loss, slowing the head increases encoder accessibility while both classifiers remain context-reliant (Table 1). Right: measured data, an illustrative validation core with its Cancer Epithelium and Cancer-Associated Stroma labels, an enlarged neighbourhood marker and processed class-mean spectra (paraffin band excised); the constructed-context experiments keep such measured centre spectra but replace the natural neighbourhood.

gradient flow on $\mathcal{L} = \mathbb{E}\log(1 + e^{-YF})$*, without other optimizer terms, from fixed* $a_0 \in \mathbb{R}$*,* $v_0 \neq 0$*,* $\sum_j \beta_{j0} = 0$*. Fix* $m > 0$ *and* $T_m = \inf\{t \geq 0 : q_J(t) \geq m\}$*,* $q_J = a + bv$*; thus* $T_m = 0$ *if* $a_0 \geq m$*.*

(a) Suppression. *If* $a_0 < m$*, set* $\Delta = m - a_0$*,* $u_M = a(T_m) - a_0$*. Then* $T_m < \infty$ *and*

$$0 < \frac{u_M}{\Delta} \leq \frac{1}{1 + Mv_0^2}, \qquad Mu_M \longrightarrow \frac{\Delta}{v_0^2} \quad (M \to \infty).$$

The spectral-only head $F = aS$ *requires update* Δ*. If* q_F *instead trains only* β *from the same initialization with* W *frozen, then* $q_J \geq q_F$ *on* $[0, T_m]$ *and* $\sup_{0 \leq t \leq T_m}(q_J - q_F) \leq C_0/(p_0 M v_0^2)$*, where* $C_0 = 1 + 2\Delta^2/v_0^2$ *and* $p_0 = (1 + e^{-a_0})^{-1}$*.*

(b) Reversal. *On* $S = Y, C = -Y$*, the stopped margin is* $2a(T_m) - m$ *when* $a_0 < m$*, and* a_0 *otherwise. Error is one if* $a_0 < m/2$ *and* $Mv_0^2 > m/(m - 2a_0)$ *(sufficient, not necessary); it is zero for* $a_0 \geq m/2$*. The spectral-only comparator has zero error for every* a_0*. For* $(a_0, v_0) \sim \mathcal{N}(0, \sigma^2 I_2)$*,* $\sigma > 0$*, with zero-sum readouts and the same stopping rule,* $\lim_{M \to \infty} \mathbb{E}_{W_0} \mathrm{Err}_{\mathrm{rev}} = \Phi(m/(2\sigma))$*, where* Φ *is the standard normal CDF.*

(c) Normalization. *Replace* b *by* $M^{-1/2}\sum_j \gamma_j$*, with* $\sum_j \gamma_{j0} = 0$ *and the same* (a_0, v_0) *and unit rates. The aggregate* (a, v, b) *dynamics equal those at* $M = 1$ *for every width.*

Proof sketch. Dividing the flow by $\dot{a} = \varrho > 0$, $\varrho = (1 + e^{q_J})^{-1}$, gives $dv/da = b$ and $db/da = Mv$, hence with $u = a - a_0$

$$v = v_0 \cosh(\sqrt{M}u), \qquad b = \sqrt{M} v_0 \sinh(\sqrt{M}u), \qquad q_J = a_0 + u + \frac{\sqrt{M}v_0^2}{2}\sinh(2\sqrt{M}u),$$

the balance invariant of gradient flow (Du et al., 2018). The fitting equation $q_J(u_M) = m$ is strictly increasing in u and $\sinh x \geq x$ gives the update bound and its limit; a transformed-margin comparison with the frozen path gives the finite-horizon gap. The sign of the reversed margin $2a(T_m) - m$ and dominated convergence give (b). For (c), differentiating the aggregate readout gives $\dot{b} = v\varrho$, independent of M.

What the theorem says, and what it does not. Replication acts as an effective learning rate on the contextual readout. At the matched margin the spectral coefficient has moved by at most a fraction $1/(1 + Mv_0^2)$ of the update the spectral-only model needs; the coefficient itself can be informative at initialization and remain so. It is the *adaptation* that is suppressed. When the initial spectral coefficient is below half the margin and the width is large enough, $Mv_0^2 > m/(m - 2a_0)$, the fitted joint model relies on context strongly enough that contextual reversal misclassifies every example, while a spectral-only comparator trained to the same margin is correct; at small Mv_0^2 the fitted model can be correct even with $a_0 < m/2$, so the failure is a finite-width statement, not a universal one. Part (c) removes the width dependence only: the normalized system is the $M = 1$ system, which for $(a_0, v_0) = (0, 1)$ still fails under reversal for every $m > 0$. Finally, training each

contextual readout at rate κ instead of unit rate gives $\dot{b} = \kappa M v \varrho$ and the bound $1/(1 + \kappa M v_0^2)$. This motivates testing relative module rates; slowing the entire learned head in the experiments is an analogous intervention, not an instance of this flow. The toy reads the spectral coordinate through a fixed unit skip, whereas the experimental head is learned; the matched-margin result does not establish the experimental equal-update ordering of encoder accessibility or recovery. A two-site extension with nonzero initial contextual readout shows that freezing the readout can leave contextual learning through the encoder active, whereas blocking the contextual-site encoder gradient makes sufficiently slow readout learning allocate the remaining margin increase to the spectral coefficient (Corollary A.1). With both routes open, an exact matched-fit equation shows that, for a nonnegative initial contextual margin, the spectral update share is non-decreasing as the readout learning rate falls, tending to $1/(1+\gamma^2 b_0^2)$ as that rate tends to zero; anti-aligned initializations admit counterexamples (Corollary A.2). Section 5 tests an analogous intervention on measured spectra.

4 Exact-model check and synthetic intervention

Experiment 2: design. Pixels carry i.i.d. labels $y_p \in \{-1, +1\}$ on 16×16 images with $S = 256$ spectral channels, so the spectral cue cannot be denoised by averaging. The spectral cue is $\alpha y_p u$ in the pixel's own spectrum, buried in isotropic noise of unit variance, with u a fixed unit direction the encoder must discover. The contextual cue places y_p in each of the eight neighbours of p, along eight fixed orthogonal directions $v_1, \ldots, v_8 \perp u$ of amplitude $\beta = 5$ with context noise $\tau\eta$: the information about y_p sits only in the neighbourhood, because p's own v-coordinates carry its neighbours' labels. The two cues are matched in signal-to-noise at oracle accuracy 0.95, calibrated before any run (Appendix C). Readiness: through a random encoder $W \in \mathbb{R}^{12\times 256}$ the same discriminant reads the contextual cue at 0.905–0.914 and the spectral cue at 0.608–0.653 (three seeds): the random projection attenuates the isotropic-noise cue by about $\sqrt{K/S}$ but not the directional one, the analogue of (a_0 small, $v_0 = O(1)$). The model is the linear encoder followed by a ReLU convolutional head (two 3×3 convolutions with width M between them, circular padding; a 5×5 receptive field), trained by full-batch gradient descent on 256 images at $\eta = 10^{-3}$ for every parameter unless an intervention multiplies a named block's rate. Runs are compared at the first step with training loss below $L^* = 0.30$ (snapshots at 0.6–0.10).

Arms and measures. The training arms are: standard parameterization at widths $M \in \{2, 4, 8, 32, 128, 512, 2048\}$; a normalized readout ($M^{-1/2}\gamma$ with $\gamma = O(1)$ and the same function at initialization, the control of Theorem 1(c)); an uninformative-context arm (the cue carrying an independent label) at $M \in \{2, 8, 128, 512\}$; a readout multiplier and a whole-head multiplier $\kappa \in \{1, 1/16, 1/256\}$ at $M = 32$; and a frozen encoder. Three registered seeds each, two added seeds agreeing; the whole-head rate arm and its recovery were extended to ten prespecified seeds (Appendix C). At every snapshot we record the probe (a discriminant on the encoder output for clean spectral-only inputs, disjoint fit and evaluation sets, reported as the gain over its initial value), its population counterpart $h_u(W) = \|P_{\mathrm{row}(W)} u\|^2$, the encoder's displacement, and the accuracies of Section 2 plus spectral-only and context-only sets. A recovery procedure, specified before its evaluations were inspected, freezes the encoder at its L^* checkpoint and trains a fresh paired-initialization head (the entire head) on new context-random images for 20,000 steps, evaluated on the same paired test family (Table 1).

Results: suppression. In the standard-parameterization baseline every model with an informative context leaves the encoder's spectral accessibility near its initial value at L^* (probe gain 0.043 at $M = 2$, 0.003 at $M = 2048$, five-seed means), with reversal accuracy 0.12–0.13 and chance accuracy under an uninformative context; trained with an uninformative context the same model aligns its encoder before reaching the same loss (gain 0.25–0.30) and survives reversal (0.88–0.90); a frozen encoder gives 0.12–0.14 (Table 4). Predictions about width, normalized-readout reversal, energy and flatness, the readout multiplier's reversal (opposite in sign) and one seed's reversal rank failed at L^* (Appendix C).

Results: where the width and speed dependence appears. At the labelled secondary threshold 0.15, spectral learning decreases with width in every seed (probe gain 0.21 at $M = 2$ falling to 0.04 at $M = 2048$, five-seed means), and the normalized readout is less suppressed at large width (Table 4). The whole-head multiplier moves the same quantity at L^*: reducing the head's rate by

256 raises mean probe accuracy from 0.65 (ten seeds) to 0.83 (nine) (Table 1) and h_u from 0.06 to 0.36 (Table 4); the tenth seed at $\kappa = 1/256$ ended at training loss 0.30008 at the 40,000-step ceiling with no matched checkpoint (post-outcome budget extension K1: Appendix C.4). Both move in the direction of the theorem's bound, below the uninformative-context control (0.89–0.94), matched at fit, not at time: at 5,000 and 15,000 steps the fast-head encoder has the larger h_u in all ten seeds. No run at $\kappa \leq 1/4096$ reached L^* (seeds 0–2 at $1/4096$, seed 0 at $1/32768$; open markers: seed 0, Figure 2a); the readout-only multiplier acts weakly the same way (Table 4).

Results: the original predictor stays context-reliant. The rate change that raises mean probe accuracy by 0.18 raises mean reversal accuracy only from 0.118 to 0.141; at an initial accessibility of $h_u(W_0) = 0.60$ instead of 0.02, slowing instead reduces reversal accuracy from 0.229 to 0.153 and enlarges the comparator deficit (Appendix C.6). Among the tested synthetic arms, only training without an informative context removes the reliance: the classifier's use of context and the encoder's accessibility respond differently to the same intervention.

Results: what retraining recovers. Over the ten prespecified seeds a fresh head reaches mean reversal accuracy 0.59 from the random and 0.59 from the fast-head encoders (ten seeds each) and 0.76 from the slow-head ones (nine seeds; Table 1; per seed: Table 5), while the original classifiers of both arms had the same poor shifted performance (0.11–0.15). Among the nine evaluable pairs, slow-head training improved fresh-head reversal recovery by 0.170 on average (sample sd 0.060; nominal paired-t 95% CI 0.124–0.216), with positive differences in all nine (the tenth pair is the censored seed). No retraining reached the 0.10 loss target, so this concerns the specified procedure and budget, not every possible repair. In a stopping-point sensitivity, continuing fast-head joint training from loss 0.30 to 0.10 raised mean fresh-head reversal recovery from 0.591 to 0.843 (ten seeds), exceeding slow-head recovery at loss 0.30 in all nine original pairs (mean paired difference $+0.086$); the original fast-head classifier still failed under reversal (mean 0.096), so the earlier stopping point's recovery disadvantage did not persist under this continuation. At equal updates the direction reverses: after 40,000 joint updates, the original slow-arm budget, mean training loss is 0.29 for the slow arm and 0.05 for the fast arm, and a fresh head recovers more from the fast-head encoder (0.917 versus 0.778; mean paired slow-minus-fast difference -0.139, negative in all ten seeds). The slow-head advantage at matched loss therefore does not extend to either tested equal-update checkpoint (Appendix C). For the original loss-0.30 encoders the advantage persisted across the tested retraining rates and budgets to 80,000 steps (Appendix C.5).

Amendments and a nonlinear encoder. A two-layer ReLU encoder (the dated amendment record is in Appendix C) reproduces the pattern in every seed: probe gain 0.00–0.02 with informative context against 0.26–0.29 without, reversal ≤ 0.13 for the ordinary informative-context arm, a $256\times$ slower head raising the gain to 0.14–0.21, and recovery 0.65–0.70 from slow-head encoders against 0.57 (fast) and 0.58 (random) (Table 1).

5 Real centre spectra with constructed context

Data and readiness. We test contextual competition with measured centre spectra and recorded labels, adding a synthetic class-associated pattern along training-estimated principal directions in the neighbourhood (natural neighbourhoods: Appendix E). The data are QCL infrared images of a commercial breast-cancer tissue microarray (169 cores, one per donor, four tissue classes; Appendix D), fold 0 of a five-fold patient-level split; the binary task Cancer Epithelium versus Cancer-Associated Stroma on 942-channel pixel vectors; 24,000 class-balanced training centres; evaluation on validation patients (test: Appendix D). A readiness scan (Table 13) shows that with per-feature standardization a random encoder $W \in \mathbb{R}^{12\times 942}$ already reads this contrast at 0.955; PCA whitening (training patients) reduces the random-encoder probe to 0.54 on training cores (0.58 on validation patients) while a calibration discriminant stays at 0.944. We therefore run two regimes on the same data, *high* and *low initial probe accessibility* (standardized and whitened inputs). Neither is an instance of the theorem (a probe does not measure the spectral share of a random head's margin).

Construction. Each patch is a real centre with eight real donor spectra drawn without conditioning on the centre label from the same split side; donor d carries $\gamma_d(c\,s + \tau\eta_d)$ along principal coordinate d, where c is the centre's label, $s = \pm 1$ or c is replaced by an independent label, and γ_d large

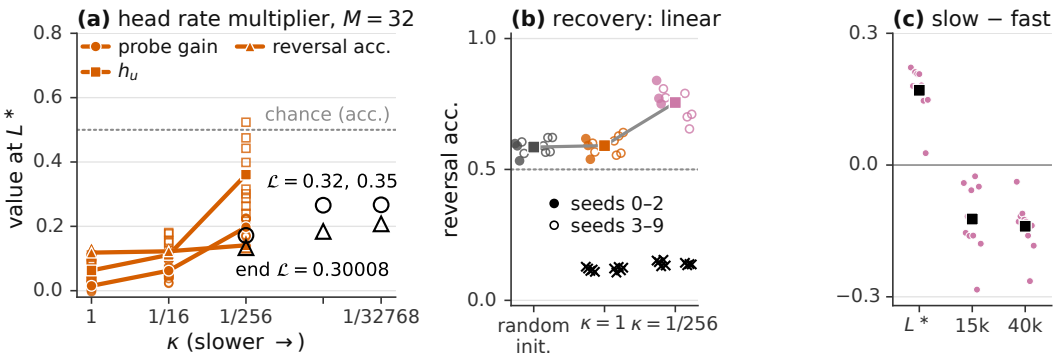

Figure 2: Experiment 2 at matched fit $L^* = 0.30$ (points: seeds, lines: means). (a) Whole-head rate κ at $M = 32$: probe gain (probe accuracy minus its own initial value), h_u and reversal accuracy over ten seeds ($n = 10$; 9 at $1/256$); open markers: budget-end states of runs that missed L^* (seed 9 at $1/256$; seed 0 at $1/4096$, $1/32768$), losses shown; dotted: chance for reversal. (b) Reversal accuracy of a fresh entire $M = 32$ head retrained 20,000 steps on context-random images from the frozen random or L^* linear encoder ($\kappa = 1$ or $1/256$): every seed (filled 0–2, open 3–9; $n = 10$, 10, 9), squares: means; crosses: original classifiers; dotted: chance. (c) Paired slow-minus-fast reversal accuracy, same protocol: encoders at matched loss (nine original pairs, mean $+0.17$) and after 15,000 and 40,000 equal joint updates (ten pairs each, -0.12, -0.14); squares: means.

relative to the donors' natural variation ($\gamma = 30$ whitened; a $\gamma = 10$ variant, added after the $\gamma = 30$ outcome, registered before its runs). τ is calibrated per regime so that a discriminant on the eight cue projections matches the centre-only one on training-side data. The model is the linear encoder over the nine spectra and a width-M ReLU head over the encoded patch, trained and thresholded as in Section 4; three registered seeds per arm, two added seeds agreeing (Appendix D).

Results: high accessibility. Values are over registered seeds 0–2 unless stated; Figure 3 and Table 13 add seeds where run. The probe is 0.91–0.94 before training and changes little at L^* in any arm, yet the informative-context heads are strongly context-reliant (reversal 0.08–0.10, chance under context-random, at every width and κ) while the comparator reaches reversal 0.93 and the frozen encoder 0.09 (Table 13, Figure 3): a large probe deficit is not necessary for failure.

Results: low accessibility. The probe starts at 0.56–0.60. Among the $\gamma = 30$ informative-context arms the mean gain at L^* is at most 0.046, against the comparator's 0.34–0.35; reversal is 0.07–0.08 against its 0.91–0.92, the frozen encoder 0.08 (Table 13). Slowing the head alone (the unblocked arms) gives only a small probe response (mean $+0.013$ from $\kappa = 1$ to $1/256$; $+0.025$ over the five seeds of Table 13 and Figure 3c): the registered $\gamma = 30$ rate prediction was not met. An intervention detaches the eight neighbours' encoder outputs from the encoder gradient (logits and head gradients unchanged at fixed parameters). At the fast head rate blocking does not consistently improve either. With the head slowed by 256×, blocking raises mean probe accuracy at matched fit from 0.63 to 0.91 and fresh-head recovery from 0.53 to 0.91, with positive differences in all three seeds. Fitting takes 43,406–65,178 steps rather than 140–226; at equal joint-update checkpoints, blocking changes slow-head recovery by -0.086 at 4,000 updates and $+0.121$ at 20,000, with the corresponding sign in every seed (Appendix D.4). The original classifiers remain context-reliant (mean reversal accuracy 0.14); preparing a head on decorrelated context over a centre-pretrained encoder raises original-classifier reversal accuracy to 0.67 at matched loss, falling to 0.14 by loss 0.10 (Appendix D.5). Thus blocking this gradient route restores a rate response in this constructed tissue setting (Table 1, Appendix D.4); neither blocking nor head preparation removed the joint classifier's reliance. On HeiPorSPECTRAL the registered primary probe and recovery benefits from blocking were not reproduced at matched loss or at 40,000 equal updates (Appendix F.2).

Results: the amplitude sweep. Across four constructed-tissue amplitudes ($\gamma = 30, 10, 3, 1$; one fold, three seeds), mean slow-minus-fast probe differences were $0.013/0.058/0.128/0.107$ and recovery differences $0.007/0.079/0.174/0.116$; the full-sweep ordering, doubling and calibrated-

Table 1: Accessibility, reliance and recovery at matched fit ($L^* = 0.30$) with a fast ($\kappa = 1$) and a slow ($\kappa = 1/256$) head at $M = 32$, means over seeds. Accessibility: linear discriminant on the frozen encoder's output for centre-spectrum-only inputs (init.: the random encoder). Recovery, rev.: reversal accuracy of a fresh entire head retrained for $20{,}000$ steps on decorrelated context from the frozen random (rand.), fast- or slow-trained encoder; procedure-dependent, not a ceiling.

		accessibility (probe)			original, rev.		recovery, rev.		
setting	n	init.	fast	slow	fast	slow	rand.	fast	slow
Synthetic, linear encoder	10/9	0.64	0.65	0.83	0.12	0.14	0.59	0.59	0.76
Synthetic, ReLU encoder	3	0.63	0.65	0.82	0.13	0.14	0.58	0.57	0.68
Tissue, high accessibility	3	0.92	0.93	0.93	0.08	0.08	0.92	0.94	0.94
Tissue, low accessibility, γ=30	3	0.58	0.61	0.63	0.07	0.07	0.50	0.52	0.53
Tissue, γ=30, neighbour route blocked[‡]	3	0.58	0.58	0.91	0.08	0.14	0.50	0.51	0.91
Tissue, low accessibility, γ=10	3	0.58	0.66	0.72	0.07	0.07	0.51	0.56	0.64

n: seeds (fast / slow). Synthetic linear: ten prespecified seeds; seed 9 at $\kappa = 1/256$ censored within the original $40{,}000$-step budget (loss 0.30008; crosses L^* at step $40{,}033$ in the registered post-outcome budget extension K1, Appendix C.4); other rows: registered seeds 0–2; tissue rows: fold 0, validation patients. Cells: binary accuracy (chance 0.50). Natural 3×3 context (Exp. 4): no reversal; its $\kappa = 1/256$ arm missed L^* within the original $40{,}000$-step budget, reached under the tenfold ceiling (Appendix E.3); budget-end informative/shuffled-context probes 0.67/0.67 (four-class macro-F1, Appendix E). [‡]Registered extension: neighbours' encoder outputs detached; slow arm reaches L^* at $43{,}406$–$65{,}178$ steps (Appendix D.4). Mean within-seed slow−fast recovery: linear $+0.17$ (9 pairs, sd 0.06), ReLU $+0.11$ (3, sd 0.03), $\gamma = 10$ $+0.08$ (3, sd 0.02), $\gamma = 30$ $+0.01$ (3, sd < 0.01), $\gamma = 30$, both arms blocked $+0.40$ (3, sd 0.01), high accessibility $+0.00$ (3, sd < 0.01). Maximum cell SD (per-cell seed SD, not pooled), in row order: at most 0.06, 0.04, 0.03, 0.07, 0.06, 0.08 (per-cell dispersions in the appendix: per-seed values, Table 5; sd bounds and ranges, Table 13's note).

peak predictions were not met; from $\gamma = 3$ to 1 the slow arm reached its best absolute outcomes but the ordinary arm improved more, narrowing the difference (Appendix D.6).

Results: what retraining recovers. Retraining separates the regimes (Table 13). With high accessibility a fresh head retrained on decorrelated context reaches reversal 0.90–0.96 from every encoder, a random one included, against 0.08 for the original classifiers: the failure coexisted with spectral information that the specified retraining could use. The nine $\gamma = 30$ encoders yield 0.45–0.56; at $\gamma = 10$ slow-head encoders yield 0.55–0.70 against 0.48–0.60 (paired gain $+0.07$ to $+0.10$).

Reading. With measured spectra, reliance matches the synthetic experiment: at every width and rate the predictor uses informative constructed context. Constructed contexts on Pavia University, Salinas and HeiPorSPECTRAL reproduce the low-accessibility suppression and reliance contrasts, while their registered $\gamma = 30$ rate predictions are not met and recovery responses depend on the setting (Appendix F). In natural 3×3 neighbourhoods, on tissue and Pavia, informative- and shuffled-context training give similar probes and comparator-level recovery (a weak comparator on Pavia, Appendix F): reliance without the constructed-context deficit, which neither explains the absence nor excludes suppression there (Appendix E). In a label-conditioned diagnostic (fold 0) unrelated same-class neighbours gave macro-F1 0.847, against 0.732 with natural and 0.234 with random-class neighbours, so disrupting the spatial arrangement alone did not reproduce the collapse.

6 Limitations and conclusion

Conclusion. In a serial linear-encoder model a faster contextual pathway suppresses the spectral coefficient's update at matched fit, and the fitted model fails under reversal under explicit finite-width conditions (Theorem 1). In the synthetic model, slowing the head raises the spectral cue's accessibility at matched loss while the classifier stays context-reliant; the fresh-head recovery contrast depends on the joint-training stopping point and the retraining protocol. With measured spectra a probe deficit is not necessary for failure and rate responses vary with the setting.

Recommendation. For the models and shifts studied here, measure contextual reliance and the linear accessibility of the centre spectral cue separately: a drop under context replacement does not show the spectrum redundant. Compare against a decorrelated-context control at a common training loss when attainable, and also at equal joint-training updates; report non-attainment. A random encoder provides an initialization baseline. Where accessibility is retained, fresh-head retraining reached comparator-level shifted accuracy in the tested settings; evaluate it first. Where

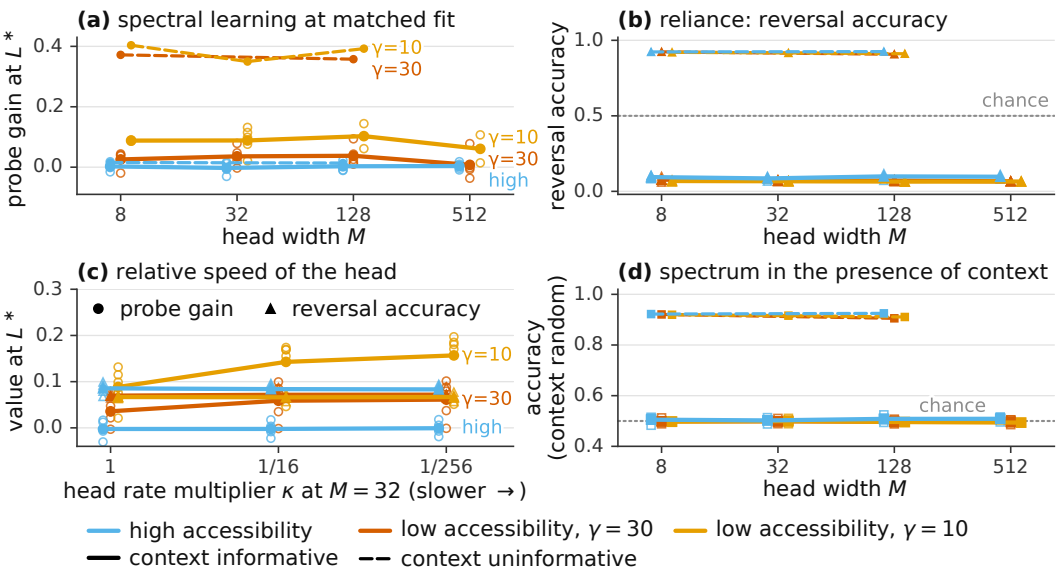

Figure 3: Experiment 3 at $L^* = 0.30$, validation patients; five seeds per primary width and rate cell, two for the added $\gamma = 10$ widths, three for its $M = 32$ comparator. Hollow: seeds; filled and lines: means; dotted: chance (b, d). Regimes: labelled in (a), (c); coincident in (b), (d), drawn left to right as high, $\gamma = 30$, $\gamma = 10$. (a) Probe gain. (b) Reversal accuracy. (c) Head rate at $M = 32$: probe gain (circles), reversal (triangles). (d) Context-random accuracy.

deficient, slowing the head improved matched-loss recovery in the synthetic model and one constructed tissue regime (reversed at equal updates), but varied across real-spectrum settings and did not repair reliance ($1/256$ is a proof of concept, not a transferable default); centre-only pretraining and neighbour-gradient blocking with a slow head improved matched-loss recoverability in the breast-QCL construction without curing reliance, and the blocking benefit was not reproduced on HeiPorSPECTRAL (Appendix F.2); we offer no decision rule.

In-distribution cost. In the natural-context experiments, decorrelated-context head retraining reduces in-distribution macro-F1 from 0.732 to 0.673 on tissue and from 0.840 to 0.637 on Pavia.

Limitations. The theorem has a linear encoder, noiseless cues and a fixed spectral skip coefficient; the experiments add noisy cues and learned routing but keep a linear encoder in the primary runs and one head family each. The low-accessibility tissue regime uses whitened inputs, the imposed context shifts are diagnostic stress tests, and transfer to minibatch adaptive optimization or deployment shifts remains untested. The cues that show the mechanism are constructed; joint-training effects of larger natural neighbourhoods remain untested, as are adaptive gradient-modulation regularizers (Peng et al., 2022). A canonical gradient-starvation remedy, Spectral Decoupling (Pezeshki et al., 2021), did not meet its accessibility, reversal or recovery criteria at the two penalty strengths reaching matched fit; the third strength was censored (Appendix C). Amplitude sweeps did not support the proposed monotone ordering of the rate benefit by initialization-time neighbour-route share, and the remedy-selection threshold failed (Appendices C.7 and D.6). Accessibility is a linear probe at one site, the encoder output; it can miss a cue that is not linearly decodable, and probe sites differ in what they support (Uselis & Oh, 2025), so the natural-context result is reliance without a detected *linear* deficit, not proof of an unimpaired encoder. Matched mean training loss is an analogue of the theorem's matched margin, not the same rule. Real-spectrum variation is measured across five patient partitions of one cohort (shared training cores: no independent-cohort generalization): the unblocked low-accessibility contrasts of Section 5 keep their sign in all five and the blocking and pretraining interventions meet their criteria (at least four of five) in all five (Appendices D.4 and D.5), while two tolerance comparisons in the natural-context experiment vary by partition (Appendix E); no general mitigation is established for natural context.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure: proposing and refining hypotheses; helping develop the theoretical model and conceptual framework; formulating mathematical claims, providing ingredients for their proofs and assisting in writing the proofs; designing the synthetic-data generators and generating the synthetic data sets; designing experiments and giving feedback on the methodology; implementing methods, including the preprocessing that cleans and reformats the data sets and the experimental code; and interpreting results. We have not used generative AI tools for translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: identifying, summarizing and analysing literature; brainstorming; suggesting experimental parameters; creating and editing software code and scientific figures; drafting, structuring and editing parts of the paper; formatting references; and adversarial review of drafts and analyses. We have reviewed all AI-assisted work: every number reported in the paper was recomputed from the recorded artifacts (the table generators and audit scripts are released), the experiments were registered before their runs and read against the registered criteria, and the theorem's predictions were checked against numerical integration of the exact model (Appendix B). We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Code, configurations and audit scripts, with seeds and exact commands, are provided in the anonymized repository accompanying this submission, which is also where archived material referenced below is kept. Every number in the paper traces to a named results file; Appendices B and H tabulate the verification checks for this paper's statements and their outcomes (the repository's README maps each to its script and command), and Appendix H names the two legacy audits of an earlier, retired framework that do not exit cleanly, with their diagnoses. Complete hypotheses and proofs are in Appendix A, with the numerical verification in Appendix B. Generators, calibration, architectures, data, patient splits and training protocols are described in Appendices C, D, E and F; Appendix G is a scope summary of a production model, whose full report is archived in the repository. The processed breast-tissue spectral images are not included in the present supplement. We plan to release these images, their annotations and patient-split files in a public data repository. The index at the head of the appendix (Table 2) maps every completed batch to its appendix, outcome and main-text location.

REFERENCES

Melih Barsbey, Lucas Prieto, Stefanos Zafeiriou, and Tolga Birdal. Large learning rates simultaneously achieve robustness to spurious correlations and compressibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2507.17748. arXiv:2507.17748.

Sebastian Berisha, Mahsa Lotfollahi, Jahandar Jahanipour, Ilker Gurcan, Michael Walsh, Rohit Bhargava, Hien Van Nguyen, and David Mayerich. Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks. *Analyst*, 144(5):1642–1653, 2019. doi: 10.1039/C8AN01495G.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023. URL https://jmlr.org/papers/v24/22-1395.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Collin Coil and Nick Cheney. What makes freezing layers in deep neural networks effective? A linear separability perspective. In *AutoML Conference*, volume 293 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of*

the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pp. 8632–8656, 2023.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1806.00900.

Ehsan Faghih, Fatemeh Ashrafi, Marguerite Moore, and Zahra Saki. Ten architectures, one error: Shared failure modes in hyperspectral classification under spatially disjoint evaluation. *arXiv preprint arXiv:2609.01786*, 2026. URL https://arxiv.org/abs/2609.01786.

Dougal Ferguson, Niels Kroeger-Lui, Domenic Dreisbach, Claire A. Hart, Diego F. Sanchez, Pedro Oliveira, Mick Brown, Noel Clarke, Ashwin Sachdeva, and Peter Gardner. Full fingerprint hyperspectral imaging of prostate cancer tissue microarrays within clinical timeframes using quantum cascade laser microscopy. *Analyst*, 150(9):1741–1753, 2025. doi: 10.1039/D5AN00046G.

Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, 2024. arXiv:2310.16228.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of Fisher information in deep neural networks: Mean field approach. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1032–1041. PMLR, 2019. URL https://proceedings.mlr.press/v89/karakida19a.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.10054. arXiv:2202.10054.

Nahal Mirzaie, Alireza Alipanah, Ali Abbasi, Amirmahdi Farzane, Hossein Jafarinia, Erfan Sobhaei, Mahdi Ghaznavi, Amir Najafi, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Implicit regularization of SGD reduces shortcut learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CPdAB7H8mU.

Edward Moroshko, Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2007.06738.

Dajana Müller, David Schuhmacher, Stephanie Schörner, Frederik Großerueschkamp, Iris Tischoff, Andrea Tannapfel, Anke Reinacher-Schick, Klaus Gerwert, and Axel Mosig. Dimensionality reduction for deep learning in infrared microscopy: a comparative computational survey. *Analyst*, 148(20):5022–5032, 2023. doi: 10.1039/d3an00166k.

Lyra O'Leary, Dougal Ferguson, Claire Hart, Mick Brown, Pedro Oliveira, Noel Clarke, Ashwin Sachdeva, Peter Gardner, and Hujun Yin. Spatial-spectral deep learning for prostate cancer tissue classification in infrared spectroscopy. *Analytical Chemistry*, 98(4):2743–2755, 2026. doi: 10.1021/acs.analchem.5c04765.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2011.09468.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2006.07710. arXiv:2006.07710.

Alexander Studier-Fischer, Silvia Seidlitz, Jan Sellner, Marc Bressan, Berkin Özdemir, Leonardo Ayala, Jan Odenthal, Samuel Knoedler, Karl-Friedrich Kowalewski, Caelan Max Haney, Gabriel Salg, Maximilian Dietrich, Hannes Kenngott, Ines Gockel, Thilo Hackert, Beat Peter Müller-Stich, Lena Maier-Hein, and Felix Nickel. HeiPorSPECTRAL – the Heidelberg porcine HyperSPECTRAL imaging dataset of 20 physiological organs. *Scientific Data*, 10:414, 2023. doi: 10.1038/s41597-023-02315-8.

Jiayi Tang, Alex Henderson, and Peter Gardner. Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. *Analyst*, 146(19):5880–5891, 2021. doi: 10.1039/D0AN02155E.

Arnas Uselis and Seong Joon Oh. Intermediate layer classifiers for OOD generalization. In *International Conference on Learning Representations*, 2025. arXiv:2504.05461.

Twan van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

Yizhou Xu and Liu Ziyin. Three mechanisms of feature learning in a linear network. In *International Conference on Learning Representations*, 2025. arXiv:2401.07085.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*. PMLR, 2021.

Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8968–8990, 2023. URL https://proceedings.mlr.press/v206/ye23a.html.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02501.

APPENDIX INDEX

Navigation, not results. Rows are grouped by scientific question; each points to the appendix holding that batch's registered criteria, populations, clause-level verdicts and censoring, which are not restated here. "Mixed" means the clauses do not share one verdict. §: main-text section (— if appendix only).

Table 2: Outcome map of the completed batches. Archived material is mapped in `docs/appendix_archive/README.md`; the previous full verdict table is `docs/appendix_archive/Index_*.tex`.

Batch (registered predictions)	Appendix	Status	§
Synthetic model: suppression, reliance and recovery			
Exp. 2 predictions P1–P6	C	mixed; P5 met	4
Nonlinear encoder, P12–P15	C.3	met	4
C6 attention (ViT-style) head, P24–P26	—	registered (pre-registration file, §21), not launched before submission: reserve behind C4 and C8; no verdict	—
Ten-seed extension	C	complete; seed 9 censored	4
K1: seed 9 under a $10\times$ ceiling	C.4	post-outcome budget extension; primary on nine pairs	4
Stopping rule, training duration and a regularizer			
K2 fast arm's later encoders, P43–P44	C.4	met	4, 6
K11 equal-update control, P78–P80	C.4, Tab. 7	not met; P78, P79 reversed	4
C10 recovery sensitivity, P30–P32	C.5	met; original fixed encoders	4
K5 Spectral Decoupling, P52–P54	C.8	not met; λ=0.1 censored	6
Rate response and its proposed predictors			
C3 denser rate grid, P18	C.6	met on uncensored rates; P18b not run	—
C2/C2b accessibility corner, P19a–c	C.6	mixed; P19b not met	4
C11/C11b amplitude, P33–P35, P39, P40, P42	C.7	mixed; P33, P34 refuted	6
C12 lower amplitudes, P36–P38, P41	D.6	mixed; only control P38 met	5, 6
C8c route share (retrospective)	D.7	not met	—
Measured tissue: constructed context and patient folds			
Exp. 3 three regimes	D	mixed; rate response not met	5
Five patient folds (Exp. 3 and 4)	D, E	mixed: sign $5/5$; P8 cl. 2 $3/5$; P11 $4/5$	5, 6
C5 pretrained encoder, P20–P23	D.5	met; reliance not cured	5, 6
Route blocking			
C8/C8b route block, P16, P17	D.4	not evaluable at original; met under a post-censoring $10\times$ ceiling	5, 6
K3 fixed-step control, P45–P48	D.4, Tab. 15	mixed by training stage	—
K8c five partitions for C8b, C5, P61–P66	D.4, D.5, Tab. 14	met in $5/5$ folds	6
Natural context			
Exp. 4 natural context, P7–P11	E	mixed, per copy	5, 6
K4 tenfold ceiling, P49–P51	E, Tab. 19	mixed; P51 not evaluable	—
K7 class-matched context, P58–P60	E, Tab. 19	mixed; P59 not met	5
C4 15×15 ring screen	E.4	primary gate failed; not launched	—
Public datasets			
Pavia University, constructed context	F	mixed; rate predictions not met	5
Pavia University, natural context	F	mixed; gate fails	1, 5
K6 per-pixel retraining, P55–P57	F	met on tissue; mixed on Pavia	—
Salinas (K10), constructed, P67–P72	F.1	mixed; rate response not met	5
Salinas (K10), natural, P73–P77	F.1	mixed; gate fails; arms censored	5
HeiPorSPECTRAL extension, HP1–HP9	F.2	mixed; rate response and matched-loss block not met	5, 6
Theory checks and scope			
Theorem 1 and Appendix A	B, H	all pass (agreement only)	3
Production model	G	freezing unresolved; drift ordering not met	1

A PROOFS OF THEOREM 1

Theorem 1 is proved once, directly, at the general initialization of its statement; the special initialization $a_0 = 0$, $v_0 = 1$ is a specialization of that proof. Three consequences follow. Notation: $G_{\boldsymbol{\theta\theta}}$, $G_{\boldsymbol{\phi\phi}}$ are the Gauss–Newton blocks of the scalar-logit loss in the blocks $\boldsymbol{\theta} = (a, v)$, $\boldsymbol{\phi} = \beta$. Auxiliary material removed in compression is archived (`docs/appendix_archive/A_*.tex`).

Proof of Theorem 1. Reduction. The training margin YF takes the same value $q = a + bv$ on both examples, so $\mathcal{L} = \log(1 + e^{-q})$ and unit Euclidean gradient flow gives $\dot{a} = \varrho$, $\dot{v} = b\varrho$, $\dot{\beta}_j = v\varrho$ with $\varrho = (1 + e^q)^{-1} \in (0, 1)$, hence $\dot{b} = Mv\varrho$. Every difference $\beta_i - \beta_j$ is conserved, so only the aggregate enters and nothing beyond $b_0 = \sum_j \beta_{j0} = 0$ is required of the initial readouts. If $a_0 \geq m$ then $T_m = 0$ by the stated stopping convention and the model is its initialization; take $a_0 < m$ until part (b). As $\varrho > 0$, $u = a - a_0$ is a time coordinate, and dividing the flow by $\dot{a}$ gives $dv/du = b$, $db/du = Mv$, so, using $v_0 \neq 0$,

$$v = v_0 \cosh(\sqrt{M}u), \quad b = \sqrt{M} v_0 \sinh(\sqrt{M}u), \quad q = a_0 + u + \frac{\sqrt{M} v_0^2}{2} \sinh(2\sqrt{M}u), \qquad (1)$$

with the balance invariant $b^2 = M(v^2 - v_0^2)$ of gradient flow (Du et al., 2018); thus v keeps its sign, $|v| \geq |v_0| > 0$ and $bv \geq 0$ for $u \geq 0$. As $\dot{a} \leq 1$, a stays finite on every finite time interval, and (1) then gives global existence.

(a) Suppression. Write $\Delta = m - a_0 > 0$ and $Q(u) = q(u) - a_0$ from (1), so that fitting means $Q(u) = \Delta$. Since $dq/du = 1 + Mv^2 + b^2 = 1 + Mv_0^2 + 2b^2 \geq 1 + Mv_0^2$, Q increases strictly from $Q(0) = 0$ to $+\infty$ and has a unique root $u_M \in (0, \Delta]$; along the flow $\dot{q} \geq \varrho \geq (1 + e^m)^{-1} > 0$ while $q \leq m$, so $T_m < \infty$ and $a(T_m) = a_0 + u_M$. As $\sinh x \geq x$, $Q(u) \geq (1 + Mv_0^2)u$, so $0 < u_M/\Delta \leq 1/(1 + Mv_0^2)$; this forces $\sqrt{M}u_M \to 0$, and expanding $\sinh$ in $Q(u_M) = \Delta$ gives $\Delta = (1 + Mv_0^2)u_M + O(M^2 u_M^3)$, hence $Mu_M \to \Delta/v_0^2$. At each fixed $u > 0$, Q increases strictly with M, so u_M decreases strictly with M. The spectral-only head $F = aS$ has $q = a$ and reaches the same margin only by moving a through all of Δ.

For the frozen comparator, hold W at (a_0, v_0) and train only β from the same initialization, and put $\Psi_{a_0}(q) = q - a_0 + e^q - e^{a_0}$, so that $\Psi'_{a_0} = 1/\varrho$. Then $\frac{d}{dt}\Psi_{a_0}(q_F) = Mv_0^2$, i.e. $\Psi_{a_0}(q_F) = Mv_0^2 t$, while $\frac{d}{dt}\Psi_{a_0}(q_J) = 1 + Mv_0^2 + 2b^2$. On $[0, T_m]$, $0 \leq bv = q_J - a \leq \Delta$ and $|v| \geq |v_0|$, so $b^2 \leq \Delta^2/v_0^2$ and the difference of the two transformed margins has derivative in $[0, C_0]$ with $C_0 = 1 + 2\Delta^2/v_0^2$; hence $0 \leq \Psi_{a_0}(q_J) - \Psi_{a_0}(q_F) \leq C_0 t$ and, Ψ_{a_0} being increasing, $q_J \geq q_F$. Convexity of Ψ_{a_0} bounds the same difference below by $\Psi'_{a_0}(q_F)(q_J - q_F)$, and convexity of the exponential gives $e^{q_F} - e^{a_0} \geq e^{a_0}(q_F - a_0)$, so $Mv_0^2 t = \Psi_{a_0}(q_F) \leq (1 + e^{-a_0})(e^{q_F} - e^{a_0})$ and $\Psi'_{a_0}(q_F) = 1 + e^{q_F} \geq 1 + e^{a_0} + p_0 Mv_0^2 t$ with $p_0 = (1 + e^{-a_0})^{-1}$. Therefore, on $[0, T_m]$, $0 \leq q_J - q_F \leq C_0 t/(1 + e^{a_0} + p_0 Mv_0^2 t) \leq C_0/(p_0 Mv_0^2)$.

(b) Reversal. On $S = Y$, $C = -Y$ the stopped model outputs $F = (a - bv)Y$, so its margin is $2a(T_m) - m$ when $a_0 < m$, using $a + bv = m$ at T_m, and is $a_0 \geq m > 0$ when $T_m = 0$, where the error is therefore zero. That margin is common to both examples, so the error is zero or one according to its sign. If $a_0 \geq m/2$ then $2a(T_m) - m = 2(a_0 - m/2) + 2u_M > 0$: error zero at every width. If $a_0 < m/2$ the margin is negative exactly when $u_M < (m - 2a_0)/2$, for which $\Delta/(1 + Mv_0^2) < (m - 2a_0)/2$ suffices, i.e. $Mv_0^2 > 2(m - a_0)/(m - 2a_0) - 1 = m/(m - 2a_0)$, the stated finite-width condition, and the error is then one. A spectral-only comparator trained to the same margin has $a = m$ and no contextual term, so its reversed margin is $m > 0$: error zero for every a_0. Finally let $(a_0, v_0) \sim \mathcal{N}(0, \sigma^2 I_2)$ with zero-sum readouts. Then $v_0 \neq 0$ almost surely, so the flow and the stopping convention apply almost surely; at fixed (a_0, v_0) with $v_0 \neq 0$ the finite-width condition holds for all large M when $a_0 < m/2$, while the error is zero at every M when $a_0 > m/2$. The events $\{a_0 = m/2\}$ and $\{v_0 = 0\}$ are null, so $\mathrm{Err}_{\mathrm{rev}}(M) \to \mathbf{1}\{a_0 < m/2\}$ almost surely, and as the errors lie in $[0, 1]$ dominated convergence gives $\lim_M \mathbb{E}_{W_0} \mathrm{Err}_{\mathrm{rev}} = \Pr(a_0 < m/2) = \Phi(m/2\sigma)$.

(c) Normalization. Replace the contextual readout by $M^{-1/2} \sum_j \gamma_j$ with $\sum_j \gamma_{j0} = 0$ and unit rates. Then $\partial q/\partial \gamma_j = M^{-1/2} v$, so $\dot{\gamma}_j = M^{-1/2} v\varrho$ and $\dot{b} = M^{-1/2} \sum_j \dot{\gamma}_j = v\varrho$, independent of M: with the same (a_0, v_0) and $b_0 = 0$ the aggregate (a, v, b) system is the $M = 1$ system at every

width, and since only the aggregate enters this holds for any zero-sum initialization of the normalized readouts. Training each unnormalized readout at rate κ instead gives $\dot{b} = \kappa M v \varrho$; replacing M by κM throughout yields the rate-κ bound $1/(1 + \kappa M v_0^2)$ quoted in Section 3. □

Remark A.1 (What the solvable model carries to the experiments). Three features of the instance matter for reading Sections 4–5. (i) The spectral coordinate is read through a fixed unit skip, while the contextual path is bilinear (readout times encoder); the bound of Theorem 1 suppresses spectral adaptation even though the spectral path is the shallower of the two. (ii) One margin $q = a + bv$ ties both coefficients. Under the zero-aggregate readout initialization of Theorem 1, with each readout trained at rate κ, as κM tends to zero at fixed remaining margin, $a(T_m)$ tends to m and the contextual contribution tends to zero; the reversed margin is positive for sufficiently small κM. This limit need not hold for a nonzero initial aggregate readout: at $\gamma = b_0 = v_0 = 1$, $a_0 = 0$, $m = 2$ with frozen readouts, Corollary A.1(i) gives $u = 1/2$ and a reversed margin of -1. In the deep heads of the experiments representation and reliance decouple, the fitted classifier staying context-reliant while a fresh head recovers (Section 4); that empirical separation between encoder accessibility and classifier reliance is therefore not itself a conclusion of this toy. (iii) Blocking the contextual-site encoder gradient in Corollary A.1(ii) freezes that coordinate outright, whereas the experimental encoder is shared, so neighbour detachment blocks only the backward route and centre-derived updates still change the neighbour representation (Appendix D.4).

Specialization to $a_0 = 0$, $v_0 = 1$, $\beta_0 = 0$. Here $\Delta = m$, (1) reads $v = \cosh(\sqrt{M}a)$, $b = \sqrt{M}\sinh(\sqrt{M}a)$, $q = a + \frac{\sqrt{M}}{2}\sinh(2\sqrt{M}a)$, and $a_M \le m/(M+1)$, $M a_M \to m$, $b(T_m)v(T_m) \ge Mm/(M+1)$. With $\Psi = \Psi_0$, $\frac{d}{dt}\Psi(q) = M + 1 + 2b^2 \in [M+1, M+1+2m^2]$ before fitting, since $b \le bv \le m$ there, so integrating from $q(0) = 0$ brackets the fitting time:

$$\Psi(m)/(M+1+2m^2) \;\le\; T_m \;\le\; \Psi(m)/(M+1), \qquad \Psi(q) = q + e^q - 1. \tag{2}$$

The initial curvature disparity is exactly the replication factor: the margin gradients are $Y(1, b)$ in $\boldsymbol{\theta} = (a, v)$ and $Yv\mathbf{1}_M$ in $\boldsymbol{\phi} = \beta$ and the scalar logistic Hessian is $\varrho(1-\varrho)$, so $\lambda_{\max}(G_{\boldsymbol{\theta}\boldsymbol{\theta}}) = \varrho(1-\varrho)(1+b^2)$ and $\lambda_{\max}(G_{\boldsymbol{\phi}\boldsymbol{\phi}}) = M\varrho(1-\varrho)v^2$, equal to $1/4$ and $M/4$ at initialization. Summing instead of averaging the cross-entropy, or rescaling the logit, multiplies both blocks equally and leaves the ratio unchanged.

Corollary A.1 (Two routes for the contextual cue). *Consider the two-site logistic model of Theorem 1, replacing $C = Y$ by $C = \gamma Y$ with $\gamma > 0$, and replacing the zero-sum readout initialization by an arbitrary aggregate $b_0 = \sum_{j=1}^{M} \beta_{j0}$. Let $M \ge 1$ be an integer, let the encoder learning rate be one and each trainable readout's learning rate be $\kappa \ge 0$, and write $\lambda = \kappa M$. Fix finite $a_0, v_0, b_0 \in \mathbb{R}$, $m > 0$ and $q_0 = a_0 + \gamma b_0 v_0 < m$, where $q = a + \gamma bv$. Define $T_m = \inf\{t \ge 0 : q(t) \ge m\}$ and $u = a(T_m) - a_0$. (i) If the readouts are frozen, then $u = (m - q_0)/(1 + \gamma^2 b_0^2)$; consequently, for $b_0 \ne 0$, the spectral update supplies strictly less than the entire remaining margin deficit. (ii) If the contextual-site contribution to the encoder gradient is blocked, so that $v(t) = v_0$, while the readouts learn, then $u = (m - q_0)/(1 + \lambda\gamma^2 v_0^2)$; for fixed initialization and γ this tends to $m - q_0$ as $\lambda \downarrow 0$, the spectral update then supplying the entire remaining margin deficit. Both hitting times are finite.*

Proof. With $\varrho = (1 + e^q)^{-1} > 0$ the flow is $\dot{a} = \varrho$, $\dot{v} = \gamma b \varrho$, $\dot{b} = \lambda\gamma v \varrho$. In (i), $b \equiv b_0$ gives $v = v_0 + \gamma b_0(a - a_0)$, hence $q = q_0 + (1+\gamma^2 b_0^2)(a - a_0)$; in (ii), $v \equiv v_0$ gives $b = b_0 + \lambda\gamma v_0(a - a_0)$, hence $q = q_0 + (1 + \lambda\gamma^2 v_0^2)(a - a_0)$. In both cases $\dot{q} \ge \varrho \ge (1 + e^m)^{-1} > 0$ while $q \le m$, so $T_m < \infty$, and setting $q = m$ gives the two displacements. □

Remark A.2. With both routes open, $u = a - a_0$ is again a time coordinate, $dq/du = 1 + \gamma^2 b^2 + \lambda\gamma^2 v^2 \ge 1$, and the root of $q(u) = m$ is unique in $[0, m - q_0]$; on that compact interval the linear (v, b) system depends continuously on λ, so the share tends to $1/(1 + \gamma^2 b_0^2)$ as $\lambda \downarrow 0$. Slowing or freezing the readouts therefore leaves a share below one whenever $b_0 \ne 0$, whereas (ii) sends it to one: the two routes are not interchangeable.

Corollary A.2 (Rate dependence and initialization). *In the setting of Corollary A.1 let both routes be open. Write $\Delta = m - q_0 > 0$, $E = \gamma^2 b_0^2$, $\Lambda = \lambda\gamma^2 v_0^2$ and $u_\lambda = a(T_m) - a_0$ for the spectral update at the matched margin; for $\lambda > 0$ put $\omega = \gamma\sqrt{\lambda}$, $s_\pm = \gamma b \pm \omega v$ and $P_\pm = (\gamma b_0 \pm \omega v_0)^2/(4\omega) \ge 0$.*

At $\lambda = 0$ the readouts are frozen, ω and $P_\pm$ are undefined, and each statement is read in its stated $\lambda = 0$ form.

(a) An exact matched-fit equation, at every rate. *For $\lambda > 0$, u_λ is the unique root in $(0, \Delta]$ of*

$$\Delta \;=\; u + P_+\big(e^{2\omega u} - 1\big) + P_-\big(1 - e^{-2\omega u}\big), \tag{3}$$

an exact implicit *equation, not a closed form; at $\lambda = 0$ the right-hand side is $(1 + E)u$, so that $u_0 = \Delta/(1 + E)$. In every case $T_m \le u_\lambda(1 + e^m) < \infty$.*

(b) A sufficient sign condition. *Suppose $b_0 v_0 \ge 0$, which includes the zero-sum initialization $b_0 = 0$ of Theorem 1. Then u_λ is non-increasing in λ and*

$$\frac{u_\lambda}{\Delta} \;\le\; \frac{1}{1+E}, \qquad \max_{\lambda \ge 0} \frac{u_\lambda}{\Delta} = \frac{u_0}{\Delta} = \frac{1}{1+E}\,: \tag{4}$$

no readout rate lifts the spectral share above $1/(1+E)$, the maximum being attained at $\lambda = 0$ (and, if $b_0 = v_0 = 0$, at every rate, since then $u_\lambda = \Delta$).

(c) Without that condition the ceiling can fail, unboundedly. *If $b_0 v_0 < 0$ both conclusions of (b) can fail. Assume in addition $a_0 < m$, which $q_0 < m$ does not imply under anti-alignment, and put $d = m - a_0 > 0$. On the family $b_0 = -\sqrt{\lambda}\, v_0$ one has $s_+(0) = \gamma b_0 + \omega v_0 = 0$, so the growing mode is annihilated ($P_+ = 0$, $P_- = \omega v_0^2 =: K$), $\Delta = d + K$ and* (3) *collapses to $u_\lambda = d + K e^{-2\omega u_\lambda}$, so $d < u_\lambda \le \Delta$ and the achieved share divided by the ceiling $1/(1 + E)$ exceeds $d(1 + E)/(d + K)$. For suitable parameters, though not on the whole family, the share exceeds the ceiling: at $\gamma = \lambda = v_0 = 1$, $b_0 = -1$, $a_0 = 0$, $m = 1$ one has $d = K = 1$, $\Delta = 2$, $1/(1 + E) = 1/2$ and the ratio is $u_\lambda = 1.10886$, the root of $u = 1 + e^{-2u}$. The ratio is unbounded along the explicit sub-family $\gamma = v_0 = 1$, $\lambda = K^2$, $b_0 = -K$, $a_0 = m - K^2$ with $K > 0$ (the example is $K = m = 1$): there $d = K^2$, $E = K^2$ and $\Delta = K^2 + K$, so the ratio exceeds $(1 + K^2)K/(K + 1) \to \infty$ as $K \to \infty$, while the share itself never exceeds 1.*

Proof. (a) Since $\varrho > 0$, $u = a - a_0$ is a time coordinate and dividing the flow by $\dot a$ gives $dv/du = \gamma b$, $db/du = \lambda\gamma v$, whence $ds_\pm/du = \pm\omega s_\pm$ and $s_\pm(u) = s_\pm(0)e^{\pm\omega u}$. With $s_+^2 + s_-^2 = 2(\gamma^2 b^2 + \lambda\gamma^2 v^2)$ this gives $dq/du = 1 + \frac{1}{2}(s_+^2 + s_-^2) \ge 1$, and integrating from $u = 0$ gives $q(u) - q_0 = u + P_+(e^{2\omega u} - 1) + P_-(1 - e^{-2\omega u})$: both bracketed terms are nonnegative, so the right-hand side is $\ge u$, strictly increasing with slope ≥ 1, and unbounded, hence has exactly one root in $(0, \Delta]$. Finally $dt = du/\varrho$ with $\varrho \ge (1 + e^m)^{-1} > 0$ on the fitting interval, and as $\lambda \to 0$ at fixed u the right-hand side tends to $(1 + E)u$, recovering Corollary A.1(i).

(b) Put $B = \gamma b_0$, $V = \omega v_0$, $y = 2\omega u$ and $\Phi_\lambda(u) = q(u) - q_0$, so $E = B^2$, $\Lambda = V^2$ and BV has the sign of $b_0 v_0$. The share does not exceed $1/(1 + E)$ at (u, λ) iff $\Phi_\lambda(u) \ge (1 + E)u$, i.e. iff $Q_{\text{ceil}} = B^2(\sinh y - y) + 2BV(\cosh y - 1) + V^2 \sinh y \ge 0$; and $du_\lambda/d\lambda = -\partial_\lambda\Phi_\lambda/\Phi'_\lambda$ with $\Phi'_\lambda \ge 1$ and $\partial_\lambda\Phi_\lambda = Q_{\text{mono}}/(4\omega\lambda)$, where $Q_{\text{mono}} = B^2(y\cosh y - \sinh y) + 2BV\, y\sinh y + V^2(y\cosh y + \sinh y)$, so u_λ is non-increasing in λ exactly where $Q_{\text{mono}} \ge 0$. As $\sinh y \ge y$ and $y\cosh y \ge \sinh y$, every coefficient of both forms is nonnegative for $y \ge 0$, so $BV \ge 0$ makes both nonnegative term by term.

The family of (c) is verified in the statement itself. □

What ρ sees. The quantifier in (b) is over deficits and rates, not trajectories, so $b_0 v_0 < 0$ is a licence to fail, not a prediction of failure. At $t = 0$ the encoder gradient is $\varrho_0(1, \gamma b_0)$, so under the norm-sum convention of the experiments the initialization route share is $\rho = \gamma|b_0|/(1 + \gamma|b_0|)$ and $E = (\rho/(1 - \rho))^2$: ρ fixes E, and with it the ceiling $1/(1 + E)$ that applies under (b), but not $\operatorname{sign}(b_0 v_0)$, which decides between (b) and (c). At $\rho = 0$ the toy has $b_0 = 0$, which falls under (b); for nonzero ρ its value alone does not establish the sign condition needed for this ceiling, and a ceiling does not order the actual benefits of a rate intervention between settings.

Proposition A.1 (Amplitude dependence of the benefit of slowing the readout)**.** *In the setting of Corollary A.2 fix b_0, v_0 with $b_0 v_0 > 0$, a remaining deficit $\Delta > 0$ (with m held fixed, holding Δ fixed means $a_0(\gamma) = m - \gamma b_0 v_0 - \Delta$) and two rates $\lambda_f > \lambda_s \ge 0$, and let the contextual amplitude $\gamma > 0$ vary, so that it scales both contextual routes, $E(\gamma) = \gamma^2 b_0^2$ and $\Lambda_i(\gamma) = \lambda_i\gamma^2 v_0^2$. Let $\mathcal{B}(\gamma) = [u_{\lambda_s}(\gamma) - u_{\lambda_f}(\gamma)]/\Delta$ be the difference of the two spectral update shares at that fixed Δ. Then (i)*

$\mathcal{B}(\gamma) > 0$ for every $\gamma > 0$; (ii) $\mathcal{B}$ is continuous on $(0, \infty)$; (iii) $\mathcal{B}(\gamma) = (\lambda_f - \lambda_s)v_0^2\gamma^2 + O(\gamma^3)$ as $\gamma \downarrow 0$; and (iv) $0 < \mathcal{B}(\gamma) \le 1/(1 + \gamma^2 b_0^2) \to 0$ as $\gamma \to \infty$. Hence $\mathcal{B}$ attains a positive maximum at an interior amplitude.

Proof. (i) Under $b_0 v_0 > 0$ the form Q_{mono} of the proof of Corollary A.2 is strictly positive for $y > 0$, so u_λ is strictly decreasing in λ. (ii) Φ_λ is jointly continuous in (γ, u) and strictly increasing in u with $\Phi'_\lambda \ge 1$, so the implicit function theorem gives a continuous root. (iii) Expanding (3) with $E, \Lambda_i = O(\gamma^2)$ gives $\Phi_\lambda(u) = u + \gamma^2(b_0^2 + \lambda v_0^2)u + 2\lambda b_0 v_0 \gamma^3 u^2 + O(\gamma^4)$, hence $u_\lambda = \Delta[1 - \gamma^2(b_0^2 + \lambda v_0^2)] - 2\lambda b_0 v_0 \gamma^3 \Delta^2 + O(\gamma^4)$, and the stated expansion follows by subtraction. (iv) is (4) applied to u_{λ_s}. A positive continuous function on $(0, \infty)$ that tends to 0 at both ends attains a positive maximum. □

Remark A.3 (Scope of Proposition A.1). This is an existence statement about the toy: it does not claim that the maximiser is unique, nor that $\mathcal{B}$ is monotone on either side of it, and the closed-form maximiser of the short-fit ($\Delta \downarrow 0$) limit is not the maximiser at a finite deficit and is not used. The toy therefore supplies no ordinal law of the form "larger contextual route strength, smaller benefit from slowing the readout": it has a rising and a falling stretch, and which stretch a sweep occupies is not determined by an initialization statistic. An experimental amplitude sweep does not in general hold the remaining margin fixed, a further reason the proposition is not a quantitative prediction for one. The hypothesis $b_0 v_0 > 0$ is not established for any network in this paper, and ρ is blind to its sign. The hypothesis also excludes $b_0 = 0$, where the outcome depends on the slow arm. With $v_0 \neq 0$ and a frozen slow arm ($\lambda_s = 0$), $u_{\lambda_s} = \Delta$ at every γ while $u_{\lambda_f} \le \Delta/(1 + \lambda_f \gamma^2 v_0^2)$ decreases strictly to zero (at $b_0 = 0$ the right-hand side of (3) is $u + \frac{1}{2}\omega v_0^2 \sinh(2\omega u)$, increasing in γ at fixed $u > 0$), so $\mathcal{B}$ rises monotonically to 1 and has no interior maximum. With both rates positive both updates tend to zero at large γ, and $\mathcal{B}$ can still have an interior maximum.

Provenance. Architecture- and initialization-dependent implicit bias (Yun et al., 2021; Moroshko et al., 2020), incremental learning (Berthier, 2023), exact deep-linear dynamics (Saxe et al., 2014), conservation of layer-norm differences (Du et al., 2018) and cross-entropy feature competition (Pezeshki et al., 2021) supply the setting and the tools used above. The serial example instantiates a known optimization-metric mechanism with finite-margin formulas, an encoder-update comparison and a specified reversal distribution; replication is equivalent to a rate M on b, and normalization removes it, so the function class does not grow. This is not a claim of a new general implicit-bias mechanism, and neither finite-time analysis nor serial composition alone is a novelty claim.

B Exact-model verification

Table 3 compares each closed form and bound of Appendix A, and its identities, with a numerical integration (DOP853, `rtol` 10^{-10}, `atol` 10^{-12}, terminal event at $q = m$) of the same flow in its full $(2+M)$-parameter space, and Figure 4 shows the mechanism; this demonstrates implementation agreement on the tested cases only. The full check table, per-case tables and the earlier draft's displacement-bound comparison are archived (`docs/appendix_archive/B_*.tex`).

C Experiment 2 details

C.1 Generator, calibration, model and amendments

Images are 16×16 tori with $S = 256$ channels per pixel, i.i.d. Rademacher labels $y_p \in \{-1, +1\}$ and orthonormal directions $u, v_1, \ldots, v_8 \in \mathbb{R}^S$ drawn once per seed (QR of a Gaussian matrix). With $o_1, \ldots, o_8$ the eight 3×3 offsets, $x_p = \alpha y_p u + \beta \sum_{d \le 8}(y_{p-o_d} + \tau \eta_{p,d}) v_d + \sigma \varepsilon_p$ ($\eta_{p,d}, \varepsilon_p$ standard Gaussian): each pixel's label is written into all eight neighbours along their own directions, while p's own v-coordinates carry its neighbours' labels, independent of y_p. Constants $\alpha = 1.645$, $\beta = 5$, $\sigma = 1$, $\tau = 1.7023$, $K = 12$. The draw order is condition-independent, so a seed's five conditions share y, η and ε: *iid*; *reversed* ($y_{p-o_d} \to -y_{p-o_d}$); *context-random* (an independent label field replaces y in the context term); *spectral-only* ($\beta = 0$); *context-only* ($\alpha = 0$). There are 256 training images and 64 per test condition; the spectral probe is fitted on 64 spectral-only images and evaluated on another 64.

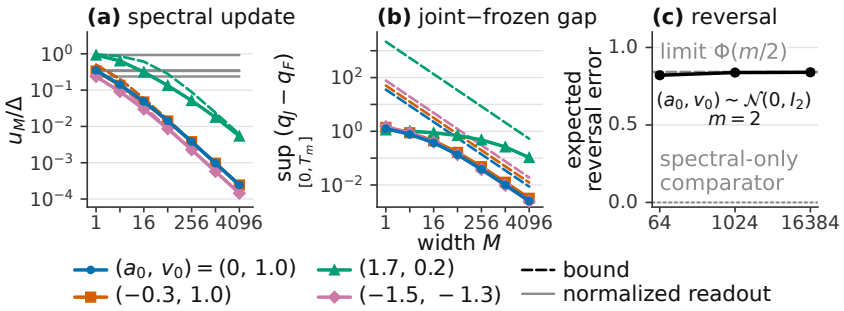

Figure 4: The exact mechanism (Theorem 1; $m = 2.9$ unless stated). (a) Spectral update fraction u_M/Δ at the matched margin against M for four initializations (a_0, v_0) (solid), the bound $1/(1 + Mv_0^2)$ (dashed) and the normalized readout of part (c) (grey, flat at the $M = 1$ value). (b) Joint-minus-frozen gap with the bound $C_0/(p_0 M v_0^2)$ (dashed). (c) Expected reversal error over $(a_0, v_0) \sim \mathcal{N}(0, I_2)$ at $m = 2$ with the limit $\Phi(m/2)$; the spectral-only comparator has error zero. Error bars in (c): three Monte Carlo standard errors ($\sqrt{p(1-p)/n}$ over the n draws), smaller than the markers.

Table 3: Checks of Appendix A, against the integrated flow except the last row (symbolic and numerical identities of the proofs): worst relative discrepancy for identities, worst slack for bounds (a negative slack would be a violation; there are none).

Statement: quantities compared	Coverage	Worst
Thm 1(a),(c): closed forms, invariant, frozen comparator, bounds	525 cases, $M \le 4096$	endpoints $\le 1.3 \times 10^{-11}$, invariant 1.52×10^{-10}; slack $+2.22 \times 10^{-10}$
Thm 1(b): reversal error at $M = 64/1024/16384$, $\sigma = 1$, $m = 2$; limit $\Phi(m/2\sigma) = 0.8413$	4,000 draws per M	$0.8207/0.8380/0.8397$
Thm 1(b): draws short of the limit that meet the width condition	4,000 per (σ, m, M) cell	0 of 48,000
Specialization $a_0 = 0$, $v_0 = 1$: endpoints, (2), $\lambda_{\boldsymbol{\phi}}(0)/\lambda_{\boldsymbol{\theta}}(0) = M$	$M \in \{16, 64, 256, 1024\}$, $m = \log 19$	1.2×10^{-13}
Cor. A.1: both identities	189 cases	1×10^{-11}
Cor. A.2: (3), (4), monotonicity in λ	540 cases	6.2×10^{-11}; exit zero requires (c)'s anti-aligned counterexamples
Prop. A.1: four clauses and an interior maximum	240 aligned; 162 (width sweep); all $b_0 > 0$, none $b_0 = 0$	240/240, 162/162 pass; uniqueness untested, unclaimed
Prop. A.1: root of (3); γ^3 coefficient	163; 240 sets	2.7×10^{-11}; 9.9×10^{-7}
Proof identities of Appendix A, symbolic and numerical	800 checks, 12 solves	All pass

Calibration (fixed before any training run), model and training. τ was set by bisection so that the contextual oracle, a Fisher discriminant on the 3×3 window of the projections $v_d^\top x$ (72 features), equals 0.950 on seed-0 directions; over seeds 0–2 it and the spectral oracle (on $u^\top x_p$) read 0.949–0.950 and 0.948–0.951. Readiness is the same discriminant through a random encoder $W \sim \mathcal{N}(0, 1/S)^{12\times256}$ (the pixel's own output for spectral-only data, the 3×3 window of outputs for context-only data): 0.608–0.653 spectral, 0.905–0.914 context (as in §4), 0.906–0.918 both. The contextual cue's population optimum, about 0.126 nats (spectral 0.127), lies just below the secondary threshold 0.15 but does not bound the finite-sample training loss, which a wide head can lower by fitting individual pixels, so losses below 0.13 are not themselves evidence of spectral learning; the probe and the shifted accuracies are. Encoder $W \in \mathbb{R}^{12\times256}$, no bias, $\mathcal{N}(0, 1/S)$ initialization; head: 3×3 convolution to width M with bias, ReLU, 3×3 convolution to one logit without bias, circular padding, PyTorch default initialization (torus-roll matrix products, equal to the convolution to 10^{-15}). Loss: mean $\log(1 + e^{-y_p F_p})$ over the 65,536 training pixels; full-batch gradient descent at $\eta = 10^{-3}$ for every parameter (the largest of $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$ with a monotone loss over 300 steps at $M = 2048$), no momentum, clipping or decay, at most 40,000 steps, stopping at loss < 0.10; snapshots at the first step below each of $0.6, 0.5, 0.4, 0.3, 0.2, 0.15, 0.10$ ($L^* = 0.30$ pre-registered). The normalized readout (`mup`) is $M^{-1/2}\gamma$ with $\gamma = \sqrt{M}\times$default (the same function at initialization; under plain gradient descent a readout-only rate $1/M$); `ctxfree` trains on context from an independent label field; widths, multipliers and rates are those of Table 4 ($\kappa = 1/4096$ on seeds 0–2 under C3 below, $\kappa = 1/32768$ on seed 0 only, §4.4c; none of these runs reached L^* within its ceiling).

Amendments that bear on interpretation (dated record: the pre-registration file released with the code). The pre-registration (§4) fixed the generator, constants, arms, L^*, measures and P1–P5 before the grid. After one seed of `sp` at $M \in \{8, 2048\}$ and `ctxfree` at $M = 8$ showed complete suppression at L^*, widths 2 and 4 and fractional readout multipliers were added and multiplier 64 dropped (§4.4a), so P1 and P2a are scored on the original and on the amended width grid separately, and P4 on the multipliers actually run; the whole-head rate arm, P6 and paired test conditions followed the design review (§4.4b); after a seed-0 pilot showed the probe but not reversal accuracy responding to κ, the κ bracket was extended below $1/256$ on seed 0 (§4.4c), an amendment prompted by an outcome of the arm it extends; the dated record (§4.4) lists every amendment with its timing. Seeds 3 and 4 were added after the registered runs (§11); a fixed policy (§19), recorded mid-batch before any slow-arm or recovery outcome of the added seeds existed (the fast arm's first L^* values were already in its log), extended the rate arm and its recovery runs to ten prespecified seeds. The nonlinear encoder's registration (§10) is the authors' record, first committed together with most of its grid (Appendix H); every extension below, from §21 on, was committed before its runs; K1, C2b and C11b were registered after the censoring each addresses had been observed and before their extension runs (§§27.1a, 21.4b, 24.2a).

C.2 Retraining (recovery) protocol, shared by Experiments 2–4

This is the single statement of the procedure; later sections state only deviations. It was specified after the original shifted results were known and before any of its own evaluations were inspected (pre-registration §6b): its design follows an external review committed before the first recovery run; the step budget and stopping rule are the authors' record, first committed with the first results (Appendix H). A frozen encoder — the random initialization or a saved trained encoder (Experiments 2 and 3: the $\kappa = 1$ and $\kappa = 1/256$ encoders saved at $L^* = 0.30$, $M = 32$; Experiment 4: the initial, L^* or final checkpoints its tables identify) — receives an entire fresh head of the original family and width ($M = 32$; every head parameter, not only a final layer), trained by full-batch plain gradient descent at $\eta = 10^{-3}$ for 20,000 steps or until training loss < 0.10 on freshly generated context-random data: 256 new images (seed $2500 + s$) in Experiment 2, new patches around the same training centres (seed $8500 + s$) in Experiment 3, context-random patches of the training centres (seed $2500 + s$) in Experiment 4, which keeps its class-weighted multiclass cross-entropy. The head initialization (seed $100 + s$) and the recovery data are identical across the encoders of a seed, so within-seed differences are paired. Evaluation is at the final state on the same paired validation or test families as the original runs, beside the original classifier's values. No completed retraining run stopped before the budget (Experiment 4 runs only the checkpoint combinations its tables list): every value is a 20,000-step endpoint, what this procedure obtains from each representation within this budget, not an information-theoretic ceiling, and the common budget does not equalize optimization difficulty across representations. A retrained head is compared with the retrained head from another encoder of the same seed or with a comparator's original classifier, labelled where both appear.

C.3 Primary results: width, rate and recovery

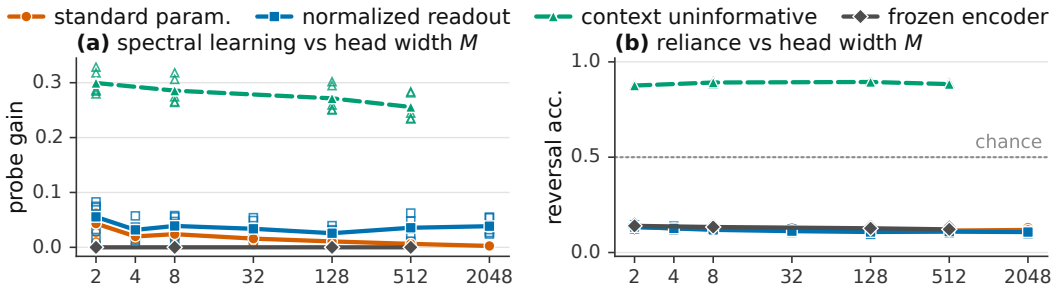

Figure 5: Experiment 2, the width sweeps at matched fit $L^* = 0.30$; points are seeds, lines means, both panels drawing the five linear seeds of the width arms. (a) Spectral probe gain against head width by arm. (b) Accuracy under context reversal.

Figure 5 gives the width sweeps (rate and recovery: main-text Figure 2); per-seed rows, per-threshold tables, unused columns and earlier notes are archived in `docs/appendix_archive/C_*.tex`. A two-layer head largely damps a readout-only

Table 4: Experiment 2 training results, mean±sd: in (a) and (b) over seeds 0–4 ($n = 5$ unless marked), in (c) over the seeds of the ten-seed rate arm that reach L^* (n shown). **(a)** Probe gain (discriminant accuracy on the encoder output minus its initial value), $h_u = \|P_{\mathrm{row}(W)}u\|^2$ and reversed accuracy at the matched fit $L^* = 0.30$, and probe gain at the labelled secondary threshold 0.15 (not pre-registered; –: not reached); context-random accuracy, used by no criterion, has cell means 0.504–0.510 (chance) in every cell of panels (a) and (b) with informative context. **(b)** Readout rate multiplier and **(c)** whole-head rate κ ($\kappa \le 1/4096$: seed-0 extension), at $M = 32$ and L^*; †: never reached L^* within the $40{,}000$-step ceiling, final state shown (loss in parentheses).

(a) width and parameterization

	`sp`, standard parameterization				`mup`, normalized readout			
M	probe gain	h_u	reversed	gain at 0.15	probe gain	h_u	reversed	gain at 0.15
2	0.043±0.024	0.083±0.040	0.133±0.007	0.213±0.030	0.056±0.030	0.097±0.049	0.134±0.008	0.226±0.025
4	0.020±0.012	0.062±0.029	0.127±0.006	0.110±0.027	0.032±0.017	0.073±0.035	0.126±0.008	0.160±0.031
8	0.024±0.012	0.063±0.020	0.123±0.002	0.120±0.019	0.039±0.021	0.077±0.025	0.119±0.002	0.150±0.031
32	0.016±0.010	0.060±0.025	0.118±0.008	0.084±0.016	0.034±0.017	0.075±0.035	0.112±0.007	0.112±0.032
128	0.011±0.005	0.054±0.021	0.116±0.010	0.061±0.016	0.026±0.012	0.067±0.025	0.108±0.009	0.089±0.025
512	0.006±0.003	0.050±0.018	0.115±0.009	0.048±0.014	0.036±0.021	0.075±0.026	0.109±0.007	0.117±0.036
2048	0.003±0.001	0.048±0.017	0.119±0.008	0.037±0.008	0.039±0.015	0.077±0.026	0.107±0.006	0.118±0.025
	`ctxfree`, uninformative context				`frozen`, encoder fixed			
M	probe gain	h_u	reversed	gain at 0.15	probe gain	h_u	reversed	gain at 0.15
2	0.299±0.022	0.858±0.035	0.876±0.002	0.309±0.023 (n=3)	0.000±0.000	0.046±0.017	0.140±0.009	–
8	0.286±0.025	0.740±0.012	0.892±0.004	0.311±0.025	0.000±0.000	0.046±0.017	0.134±0.006	–
128	0.272±0.025	0.651±0.010	0.895±0.003	0.308±0.025	0.000±0.000	0.046±0.017	0.127±0.008	–
512	0.256±0.025	0.563±0.004	0.883±0.004	0.306±0.025	0.000±0.000	0.046±0.017	0.122±0.008	–

(b) readout learning-rate multiplier, $M = 32$, 5 seeds

multiplier	0.0625	0.25	1	4	16
probe gain	0.031±0.016	0.023±0.015	0.016±0.010	0.011±0.005	0.007±0.002
h_u	0.073±0.034	0.067±0.030	0.060±0.025	0.055±0.021	0.051±0.019
reversed	0.112±0.007	0.114±0.007	0.118±0.008	0.124±0.008	0.130±0.009

(c) whole-head rate multiplier κ, $M = 32$

κ	n	step	probe acc.	probe gain	h_u	reversed	context-random
1	10	1,789±115	0.655±0.034	0.015±0.012	0.063±0.029	0.118±0.008	0.506±0.006
0.0625	10	9,019±949	0.702±0.047	0.062±0.030	0.111±0.054	0.122±0.007	0.505±0.006
0.00390625	9	33,934±5,515	0.834±0.034	0.197±0.032	0.361±0.102	0.141±0.009	0.509±0.008
0.00390625† (seed 9)	1	40000 (0.300)	0.840	0.172	0.359	0.134	0.509
0.000244141† (seed 0)	1	40000 (0.320)	0.913	0.266	0.693	0.184	0.529
3.05176e-05† (seed 0)	1	40000 (0.347)	0.914	0.267	0.701	0.208	0.528

The 3 registered runs that ended at the step budget above L^* are the rows marked † in panel (c); every other registered run reached L^*.

Table 5: Paired recovery at $M = 32$, one column per seed: reversal accuracy of a fresh entire head (Appendix C.2) from the frozen random encoder, the ordinary head's L^* encoder (fast, $\kappa = 1$) and the slow head's ($\kappa = 1/256$), their within-seed difference, and the two arms' own classifiers at L^* on the same paired test family. $*$: seed 9 has no $\kappa = 1/256$ checkpoint at L^* (censored at the $40{,}000$-step ceiling) and contributes no pair; means over the seeds with a value (10; 9 in the slow rows), the difference over 9 pairs.

seed	0	1	2	3	4	5	6	7	8	9	mean
random	0.599	0.590	0.533	0.605	0.561	0.590	0.566	0.622	0.568	0.622	0.585
fast	0.618	0.591	0.539	0.600	0.567	0.609	0.554	0.628	0.562	0.641	0.591
slow	0.840	0.771	0.750	0.808	0.774	0.791	0.701	0.655	0.711	–*	0.756
slow−fast	+0.222	+0.180	+0.212	+0.208	+0.207	+0.182	+0.146	+0.027	+0.148	–*	+0.170 (sd 0.060)
orig. fast	0.128	0.121	0.112	0.118	0.109	0.121	0.106	0.128	0.115	0.125	0.118
orig. slow	0.152	0.146	0.130	0.155	0.133	0.143	0.134	0.139	0.137	–	0.141

multiplier (Table 4(b)), its first convolution still adapting at full rate; the theorem's head is a single linear readout, of which the whole-head multiplier is the closer experimental analogue, not a proved reduction.

Seed populations and the censored slow-head seed. Seeds 0–2 are registered and decide P1–P6; the added seeds 3 and 4 enter the means of Tables 4 and 5 and confirm the suppression, rate and recovery comparisons (probe gain 0.008 and 0.005, rising to 0.225 and 0.221 under the slow head), but seed 3 has a positive width–reversal rank correlation ($+0.7$), so the registered width prediction remains unmet with five seeds as with three. In the ten-seed rate arm, seed 9 at $\kappa = 1/256$ ended at training loss 0.3000819981 at the $40{,}000$-step ceiling, with no L^* checkpoint and no matched-fit

Table 6: Registered criteria for Experiment 2 (pre-registration §4.4; the ReLU-encoder block §10) at $L^* = 0.30$, on the original five widths and on the amended grid; r_s = Spearman rank correlation across widths (or multipliers), one value per registered seed (0–2); $a_u(W) = \|Wu\|^2/\|W\|_F^2$ is the share of the encoder's squared weight mass on u.

criterion (population)	decisive statistic	verdict
P1: standard param., $r_s(a_u, M) \leq -0.8$ and $r_s(\text{rev}, M) \leq -0.8$ in every seed	original: a_u: -0.70, -1.00, -1.00; rev: -0.60, -0.90, -0.60	not met (met in 1/3 seeds)
	amended: a_u: -0.54, -1.00, -0.96; rev: -0.64, -0.96, -0.82	not met
P2a: normalized readout, $\lvert r_s(\cdot, M)\rvert < 0.5$ in at least two seeds	original: a_u: $+0.10$, -0.90, -0.30; probe: -0.20, -0.90, -0.30; rev: -0.90, -0.70, -0.70	alignment met; probe met; reversal not met
	amended: a_u: $+0.25$, -0.96, -0.54; rev: -0.89 in every seed	alignment not met; reversal not met
P2b: $a_u(2048)$ normalized > standard in every seed (paired seeds)	0.0059 vs 0.0045; 0.0062 vs 0.0045; 0.0043 vs 0.0027	met
P3: uninformative context, $a_u > 0.25$ at every width and $\lvert r_s(a_u, M)\rvert < 0.5$ (amended control widths $\{2, 8, 128, 512\}$)	a_u range 0.058–0.174; $r_s(a_u, M)$: -1.00 in every seed; $r_s(\text{probe}, M)$: -1.00 in every seed	cutoff not met; flatness not met (probe 0.936→0.893)
P4: readout multiplier, a_u and reversal accuracy decrease with the multiplier in every seed (run multipliers $\{1/16, 1/4, 1, 4, 16\}$; 64 dropped)	a_u: -1.00 in every seed; rev: $+1.00$ in every seed	a_u met; reversal not met (opposite sign)
P5: frozen encoder, reversal accuracy < 0.5 at every width (amended widths)	max 0.146	met
P6 (amendment 4.4b): whole-head $\kappa \in \{1, 1/16, 1/256\}$, three seeds: probe, h_u and reversal accuracy increase as κ decreases, $r_s \leq -0.8$ in every seed	probe: -1.00 in every seed; h_u: -1.00 in every seed; rev: -1.00, -1.00, -0.50	probe and h_u met; reversal not met (met in 2/3 seeds)
Two-layer ReLU encoder (pre-registration §10), seeds 0–2*: ranges over seeds*		
P12: informative-context probe gain below the uninformative comparator's at $M = 8$ and 128 in every seed	informative 0.001–0.022; comparator 0.256–0.286	met (every seed)
P13: informative-context reversal ≤ 0.5 at every width and seed (all arms), comparator ≥ 0.75	standard ≤ 0.128, slow head 0.131–0.154, frozen 0.139–0.153; comparator ≥ 0.879	met (every seed)
P14: probe gain at $\kappa = 1/256$ above $\kappa = 1$ ($M = 32$, the two endpoint rates) in every seed, mean reversal change < 0.05	0.142–0.210 against 0.016–0.020; reversal $+0.018$	met (every seed)
P15: fresh-head reversal (recovery protocol, $M = 32$) from the $\kappa = 1/256$ encoder above $\kappa = 1$ in every seed; $\kappa = 1$ at most 0.05 above random	0.650–0.699 against 0.566–0.572 (random 0.579–0.580)	met (every seed)

recovery value: slow-arm attainment is $9/10$ against fast-arm $10/10$, so the censoring falls in the treatment arm. The primary analysis stays on the nine pairs evaluable within the prespecified budget: paired slow-minus-fast recovery difference $+0.170$ (sd 0.060, positive in $9/9$, nominal paired-t 95% CI 0.124–0.216), $+0.205$ over the three registered seeds and $+0.153$ over the six added evaluable ones. Any slow-arm value in $[0, 1]$ for the missing seed (fast arm 0.641) leaves the hypothetical ten-seed mean difference between 0.089 and 0.189 — a bounded-completion sensitivity analysis, not a confidence interval. Every recovery statement here is conditional on those nine pairs. Extending the censored run's ceiling after this outcome was known neither replaces a seed nor extends the sample (what the ten-seed policy forbids) but changes the primary protocol, so that extension (K1, Appendix C.4) is only a labelled post-outcome budget extension.

Nonlinear encoder (registration §10). A two-layer ReLU encoder ($256 \to 64 \to 12$; first layer $\mathcal{N}(0, 1/S)$ with zero bias, second $\mathcal{N}(0, 1/64)$) replaces the linear one, all else unchanged, seeds 0–2: informative context at $M \in \{8, 32, 128, 512\}$, uninformative at $M \in \{8, 128\}$, $\kappa = 1/256$ and a frozen encoder at $M = 32$; all twenty-four runs reached L^*. With a_u and h_u undefined, the probe, the paired shifted accuracies and recovery carry P12–P15, all met in every seed (Table 6); the random ReLU encoder reads the spectral cue at 0.635, the linear encoder's readiness. Recovered levels are lower than with the linear encoder (slow-head 0.75–0.84 there, seeds 0–2), in the same order.

C.4 Stopping rule and training duration (K1, K2, K11)

The rate arms are compared at the first crossing of L^* (1,580–1,977 steps at $\kappa = 1$, 22,708–39,326 at $\kappa = 1/256$). At equal steps the ordering reverses: at exactly the same logged step of the recorded trajectories (no new run) the $\kappa = 1$ encoder has the larger h_u in all ten prespecified seeds at 1,000, 2,000, 5,000, 10,000 and 15,000 steps (means 0.10 and 0.31 at 5,000 and 15,000 against 0.07 and

Table 7: Stopping rule and training duration ($M = 32$): three registered checks, with populations and registration dates as labelled; recovery is the fresh-head reversal accuracy (Appendix C.2). K11's registered table is panel (c), merged here by the presentation amendment of 24 September.

(a) K1, post-outcome budget extension (registered 21 September, a week after the censoring was recorded): the censored slow seed under a 400,000-step ceiling.

	first crossing of L^* (step)	probe at L^*	original reversal	recovery (reversal acc.)	slow−fast recovery difference
seed 9, 400,000-step ceiling	40,033	0.840	0.133	0.742	+0.101 (fast seed 9: 0.641)
registered primary, nine pairs, unchanged	–	0.834	–	0.756	+0.170 (sd 0.060; 9/9 positive)
ten-pair secondaries† S3, S2, S1	–	0.835	–	0.754	+0.163 (sd 0.061; 10/10 positive)

Seed 9 crossed L^* at step 40,033 and ran on to the stop rule (loss < 0.10) at step 366,314; its recovery is paired with the registered $\kappa = 1$ and random-encoder rows of seed 9. Population rows: means over the nine pairs evaluable within the prespecified 40,000-step budget (the registered primary estimate) and over seeds 0–9 (the ten-pair secondaries S3, S2, S1, defined before the extended run).
† Ten-pair secondary, includes one run extended to a 400,000-step ceiling after its censoring was observed; reported beside the primary estimate, never in its place, and in no figure, table or verdict whose registered population is the nine pairs.

(b) K2, sensitivity, not a primary endpoint (registered 21 September, after the matched-fit result): fast encoders at joint loss 0.30, 0.15, 0.10; matched on neither steps nor loss.

encoder saved at	joint steps	n	probe	recovery	recovery − fast L^*	recovery − slow L^*	own classifier
random initialization	0	10	0.640±0.025	0.585±0.028	−0.005 (3/10)	−0.174 (0/9)	–
fast, loss 0.30 (L^*)†	1,789	10	0.655±0.034	0.591±0.034	–	−0.170 (0/9)	0.118
fast, loss 0.15	7,883	10	0.725±0.048	0.660±0.059	+0.069 (10/10)	−0.102 (1/9)	0.072
fast, loss 0.10	20,607	10	0.867±0.011	0.843±0.010	+0.252 (10/10)	+0.086 (9/9)	0.096
slow, loss 0.30 (L^*)†	33,934	9	0.834±0.034	0.756±0.058	+0.170 (9/9)	–	0.141

P43 (seeds 0–9; loss-0.10 fast minus fast L^* at least +0.05 on average and positive in at least 8): +0.252, 10/10 positive — *met*
P44 (seeds 0–8; loss-0.10 fast minus slow L^* at least −0.03): +0.086 (sd 0.058), 9/9 positive — *met*

Encoders saved at the first step with joint loss below the stated value (fast: $\kappa = 1$; slow: $\kappa = 1/256$, seeds 0–8); † the registered encoders of Table 5, whose rows, with the random row, this invocation reproduces exactly (maximum absolute difference 0); all 49 retraining runs ended at the step budget. Differences: paired within seed, mean (positive seeds / pairs); own classifier: the joint run's own classifier under reversal. Recovery rises from L^* through loss 0.15 to 0.10 in 10 of 10 seeds (reported, not predicted).

(c) K11, equal-update control (registered 24 September): both arms after exactly 15,000 and 40,000 joint updates; matched on fit at neither.

head rate	updates	training CE	joint model, reversed	probe	fresh head, reversed	n
$\kappa = 1$	15,000	0.119 (0.10–0.14)	0.081	0.808	0.768±0.082	10
$\kappa = 1/256$	15,000	0.379 (0.34–0.40)	0.168	0.745	0.645±0.061	10
$\kappa = 1$	40,000	0.045 (0.04–0.06)	0.264	0.930	0.917±0.011	10
$\kappa = 1/256$	40,000	0.286 (0.25–0.30)	0.136	0.850	0.778±0.058	10

P78 (primary), fresh-head recovery at 40,000 updates, slow minus fast: −0.139, negative in 10/10 seeds — *reversed*
P79, fresh-head recovery at 15,000 updates: −0.123, negative in 10/10 — *reversed*
P80, probe at 40,000 updates: −0.080, negative in 10/10 — *not met*

Training CE: achieved training cross-entropy at the checkpoint, mean (range); the registered matched-fit level is $L^* = 0.30$. Fresh head: mean±sample sd over the ten prespecified seeds. Verdicts on the within-seed paired differences: met if the mean is at least +0.05 with the sign in at least eight of ten seeds, reversed if at most −0.05 with the opposite sign in at least eight, otherwise not met.

0.17) and in the four seeds whose $\kappa = 1$ run is still training at 20,000 — at these checkpoints, not at every step (one seed's $\kappa = 1$ encoder falls slightly below between steps 10,550 and 13,525); trained on to loss 0.10, the $\kappa = 1$ encoder reaches probe gain 0.23 (five seeds). *K1* changed only the censored run's ceiling and, before any outcome was read, reproduced that run exactly on every shared trajectory row, snapshot and encoder array (maximum absolute difference 0); it adds no seed, replaces none and changes no analysis population. *K2* repeats no joint run; its loss-0.10 fast encoder has had fewer joint steps than the slow encoder and a lower joint loss (no step-matched control), yet it supports a fresh head at 0.843, above the slow encoder's 0.756 at L^* in all nine original pairs: the recovery disadvantage at the earlier stopping point did not persist under this continuation, while the fast classifier stays context-reliant (reversal accuracy 0.096). *K11* reran both arms for exactly 40,000 updates with the loss stop disabled (identity gate: the twenty runs reproduce the original trajectories exactly up to their stopping points); both recovery predictions are reversed, at checkpoints where the fast arm has also reached a far lower training loss, and the probe prediction is not met. The rate result is therefore greater accessibility at the matched-loss stopping point; the slow-head advantage does not extend to either tested equal-update checkpoint, identifies no improvement per update or per unit of compute, and leaves the matched-fit comparison of Table 5 unchanged.

Table 8: C10: slow−fast recovery on the original L^* encoders by recovery rate η, budget (reads of one 80,000-step trajectory per η) and standardization.

		5,000 steps			20,000 steps			80,000 steps		
η	features	slow−fast (sd)	pos.	|f−r|	slow−fast (sd)	pos.	|f−r|	slow−fast (sd)	pos.	|f−r|
10^{-2}	raw	+0.18 (0.05)	9/9	0.01	+0.19 (0.05)	9/9	0.01	+0.20 (0.05)	9/9	0.01
10^{-3}	raw	+0.11 (0.05)	9/9	0.00	+0.17 (0.06)	9/9	0.01	+0.19 (0.05)	9/9	0.01
10^{-4}	raw	+0.04 (0.03)	8/9	0.00	+0.09 (0.04)	9/9	0.00	+0.12 (0.06)	9/9	0.00
10^{-3}	standardized	+0.04 (0.03)	8/9	0.01	+0.08 (0.06)	8/9	0.00	+0.18 (0.06)	9/9	0.01

slow−fast: mean within-seed paired difference in fresh-head reversal accuracy over the nine eligible pairs (sample sd); pos.: pairs with a positive difference, of the fixed nine; |f−r|: absolute mean fast−random difference over ten seeds; standardized: the head sees $(z - \mu)/\sigma$ per channel with training-image statistics.

C.5 Recovery-stage sensitivity: rate, budget and standardization (C10)

Registered before its runs and amended before any outcome was read (§22.3, §22A): the recovery protocol rerun on the saved random, $\kappa = 1$ and $\kappa = 1/256$ L^* encoders of the original ten-seed rate arm (slow seed 9 censored) at three learning rates, each along one trajectory with an 80,000-step ceiling and the registered stop rule, read at three budgets, plus one $\eta = 10^{-3}$ arm on per-channel standardized features. *P30* (every unstandardized cell: slow-minus-fast positive in at least eight of the nine eligible pairs, fast-head encoders within 0.03 of random)*: met*, nine of nine in eight cells, eight of nine at $\eta = 10^{-4}$/5,000 steps, $|\text{fast} - \text{random}| \leq 0.010$ throughout. *P31* ($\eta = 10^{-3}$ gain at 80,000 steps at least half that at 20,000)*: met*, +0.187 against +0.170. *P32* (standardization keeps at least half the unstandardized 80,000-step gain)*: met*, +0.183. No trajectory diverged or reached the stop loss; the $\eta = 10^{-3}$, 20,000-step cell reproduces the registered recovery exactly. C10 varies only the recovery stage for the original loss-0.30 encoders: persistence within this protocol family, not optimal recoverability; standardization tests marginal feature scales only; it does not test equal joint-training time (Table 7(c) does).

C.6 Head-rate grid and initial accessibility (C3, C2, C2b)

Table 9: C3: the seven whole-head rates at matched fit ($M = 32$, three-seed means). †: censored, not all seeds reached L^* in 40,000 steps; budget-end means.

κ	1	1/4	1/16	1/64	1/256	1/1024 †	1/4096 †
seeds at L^*	3/3	3/3	3/3	3/3	3/3	1/3	0/3
probe gain	0.022	0.046	0.087	0.145	0.210	0.237	0.239
h_u	0.067	0.090	0.141	0.240	0.406	0.518	0.524
reversal	0.121	0.120	0.123	0.132	0.143	0.162	0.177

Table 10: C2 and C2b: the four corner cells of initial spectral accessibility h_0 against head rate κ ($M = 32$, seeds 0–2, matched fit).

cell (h_0, κ)	A_{rev} informative	A_{rev} twin	mean D	D per seed	status
$(0.02,\ 1)$	0.117	0.897	+0.779	+0.772, +0.782, +0.783	FAILURE
$(0.02,\ 1/256)^\dagger$	0.142	0.926	+0.784	+0.770, +0.784, +0.798	FAILURE
$(0.6,\ 1)$	0.229	0.908	+0.679	+0.674, +0.683, +0.680	FAILURE
$(0.6,\ 1/256)^\dagger$	0.153	0.925	+0.772	+0.757, +0.767, +0.791	FAILURE

The informative arm is `c2`, its uninformative-context twin `c2_ctxfree`; $D = A_{\text{rev}}(\text{twin}) - A_{\text{rev}}(\text{informative})$ is the reversal deficit. A cell is ELIGIBLE only when both arms reach L^* in all three seeds and the twin has mean reversal ≥ 0.85 with no seed below 0.80; FAILURE is mean $D \geq 0.10$ with $D \geq 0.10$ in at least two seeds, RECOVERED is mean $D \leq 0.03$ with $D \leq 0.05$ in every seed. Accuracies are three-seed means; D is paired by seed; the comparison is at matched fit, not matched time.

†Twins from the extension C2b (plan §21.4b, registered after the censored corner outcome was read and before its runs): the six censored runs re-run from scratch with a 400,000-step ceiling, nothing else changed; before any number below was read each reproduced its censored counterpart exactly at every shared point (1,600 trajectory rows, 11 threshold snapshots, the step-40,000 state; maximum absolute difference 0). The informative arm is not re-run: its L^* snapshot, the first crossing of 0.30, cannot move.

C3 (§21.3) extends the whole-head rate arm from three multipliers to seven, everything else unchanged. *P18* (the three-seed means of probe gain and h_u at L^* each monotone non-decreasing along increasing $1/\kappa$ over the rates whose three seeds all reach L^*, at least five such rates required)*: met over the five uncensored rates*, the registered minimum, satisfied exactly, with no margin and no tolerance (any negative step counts as a decrease); 16 of 21 cells reached L^*, and the two censored rates enter no verdict. No rate law, saturation level or functional form is predicted or claimed; P18b, the conditional extension, was not run.

C2 (§21.4, extension §21.4b) sets $h_0 = \|P_{\mathrm{row}(W_0)}u\|^2$ to 0.02 or 0.60 by an input-space rotation fixing the contextual-cue span pointwise (pre-launch identity test: deviation at most 3.9×10^{-9}, no randomness consumed), crossed with $\kappa \in \{1, 1/256\}$. P19a: mean D no larger at the higher h_0 in each κ column; P19b: no larger at the slower head in each h_0 row; each also requires that no seed move against that direction by more than 0.02 (a per-seed tolerance of our own, stricter than the mean-level tolerance of the full-grid criterion; the mean clauses carry none); P19c: the default cell $(0.02, 1)$ is a FAILURE. All six slow-head twins were censored at the 40,000-step ceiling, leaving both slow-head cells ineligible and both column predictions unevaluable; C2b, registered after this censoring was read and before its runs, re-ran exactly those six at 400,000 steps, all reaching L^* at 1.9–4.3 times the old ceiling. All four cells are then eligible and all four are a FAILURE: the classifier stays context-reliant at matched fit in every corner. *P19a: met*, by small margins: mean D falls by 0.100 with h_0 at $\kappa = 1$ and by 0.012 at $\kappa = 1/256$, about 1.6% of the slow-head deficit. *P19b: not met, reversed*: at $h_0 = 0.60$ the deficit is *larger* with the slower head (+0.772 against +0.679, a violation of 0.093, beyond the tolerance in all three seeds; 0.005 at $h_0 = 0.02$), so the verdict does not turn on the tolerance. *P19c: met*, restated from the default cell, not recomputed. At $h_0 = 0.60$ slowing makes the informative classifier *worse* under reversal ($0.229 \to 0.153$) and the comparator *better*, so D, a difference of two classifiers, is not an encoder-information deficit. Four cells are a discovery stage (the 5×5 grid is a separate registration): no boundary in (h_0, κ) is claimed or observed, and the toy's boundary is not drawn over this map.

C.7 AMPLITUDE SWEEP AND ROUTE SHARE (C11, C11B)

Table 11: C11 and C11b: the four-level contextual-amplitude sweep ($M = 32$, seeds 0–2, matched fit).

β	τ	$\bar{\rho}$	rate response R (per seed; mean)	recovery gain Q (per seed; mean)	$\tilde{R}$, $\tilde{Q}$
0.75‡	1.07465	0.5110	+0.041, +0.038, +0.040; +0.0398	+0.045, +0.042, +0.048; +0.0447	(3.705, 3.643)§
1.5‡	1.57709	0.6727	+0.086, +0.079, +0.103; +0.0895	+0.098, +0.087, +0.123; +0.1024	1.337, 1.231
3‡	1.68146	0.8020	+0.187, +0.178, +0.221; +0.1954	+0.215, +0.200, +0.263; +0.2257	0.957, 0.869
5†	1.70233	0.8644	+0.195, +0.178, +0.192; +0.1884	+0.222, +0.180, +0.212; +0.2045	0.740, 0.624
probe at L^*, $\kappa = 1$, per seed, $\beta = 0.75/1.5/3/5$: 0.901, 0.907, 0.904 / 0.851, 0.861, 0.834 / 0.728, 0.733, 0.674 / 0.675, 0.683, 0.628					
achievable reference A, same order: 0.912, 0.916, 0.917 / 0.912, 0.916, 0.918 / 0.913, 0.917, 0.918 / 0.914, 0.918, 0.918					

Per-seed values in seed order. $\bar{\rho}$ is the three-seed mean neighbour-route share of the encoder gradient at initialisation, measured before that level trained. R is the probe gain at $\kappa = 1/256$ minus the probe gain at $\kappa = 1$, both read at the first training-loss crossing of $L^* = 0.30$; Q is the fresh-head reversal accuracy of the $\kappa = 1/256$ encoder minus that of the $\kappa = 1$ encoder; $\tilde{R}$ and $\tilde{Q}$ normalise them by the decorrelated-context headroom; A is the decorrelated-context probe at L^*. † $\beta = 5$ is the pre-existing registered setting and a fitted anchor of the toy calibration, not a prediction.

‡ the $\kappa = 1/256$ runs at $\beta = 0.75$, 1.5 and 3 reached L^* only under the registered $10\times$ ceiling of 400,000 steps (extension C11b, registered after these censored outcomes were read and before its runs): under the original 40,000-step ceiling eight of these nine runs were censored. Every re-run reproduces the censored run exactly at all shared threshold snapshots, on the shared trajectory rows to step 40,000 and at step 40,000 (max |difference| = 0); no run is still censored at 400,000.

§ excluded from P40 as *no headroom* (denominator < 0.05); the values are reported, not scored.

Registered before its runs (§24, §24A; toy-calibrated predictions recorded before any outcome; the extension §24.2a registered after the censoring below was observed): four contextual amplitudes $\beta \in \{0.75, 1.5, 3, 5\}$, τ recalibrated per level to hold the contextual oracle at 0.95. The gate P35 passed in all three clauses, two at machine precision. Eight of the nine slow-head runs at $\beta \leq 3$ were censored at the 40,000-step ceiling; C11b, registered after that censoring was read, re-ran all nine at 400,000 steps, changed nothing else and passed its identity gate, so all four levels are eligible. *P33, P34* (R, Q non-increasing in ρ, separated by at least 0.05 between the extreme levels; refuted by one seed-level inversion beyond 0.05 for R or 0.08 for Q)*: refuted* — both raw responses rise steeply with ρ over most of the sweep (extreme-level separations -0.149 and -0.160; four seed-level inversions beyond 0.05 for R, three beyond 0.08 for Q). *P40*, the same ordering on $\tilde{R}$ and $\tilde{Q}$*: met on the three levels with headroom*; $\beta = 0.75$ is excluded as registered, its denominator falling below 0.05. The normalization divides by the gap between the ordinary ($\kappa = 1$) arm and a decorrelated-context comparator at the same matched fit, an *achievable* reference, not a ceiling, so values above one occur. P40 is a separate normalized result, not a rescue of P33 and P34, and is itself confounded: its denominator collapses at low β because the ordinary arm's probe rises towards the comparator. *P42*, calibrated from the toy and predicting the opposite ordering (so at most one of P33 and P42 can be met)*: not met* — its argmax clause fails and, at the three levels that are not fitted anchors, the responses lie 2.3–6.6 times above the calibrated envelope. The largest sample means occur at $\beta = 3$, but the probe difference between $\beta = 3$ and 5 is within the declared 0.01 tie band;

the recovery margin, 0.021, is larger, but no tie band was registered for the location of the recovery maximum. Both arms' probe accuracies fall as β rises while the comparator stays at 0.915–0.917: the response grows because the ordinary arm degrades faster, not because the slow arm improves.

P39 tests a declared, not derived, decision rule (retrain the head when initial accessibility is high; otherwise slow the head when mean $\rho < 0.85$ and train with decorrelated context when $\rho \geq 0.85$): of three complete procedures, each ending in fresh-head retraining, the predicted one must attain the highest fresh-head reversal accuracy at every eligible level in at least two of three seeds (ties within 0.02 consistent). Initial accessibility is low at every level (random-encoder probe 0.636 against the rule's 0.85), so only the second branch is exercised. *P39: not met* — at $\beta = 3$ the predicted remedy is best in no seed and near-best in one; the crossover between the two remedies lies below the cut, between $\rho = 0.673$ and 0.802. The rule is best in $9/12$ eligible level×seed cells against at most $6/12$ for any constant policy, so its registered no-selection clause does not fire; that does not rescue the failed clause, and no prescriptive rule for choosing remedies follows. With Appendix D.6, these sweeps did not support ordering the rate benefit by the initialization-time route share, and the toy-calibrated peak predictions were met in neither: that share is not the right summary of the route split here, though the route explanation may still hold dynamically (Appendix D.4). Amplitude also changes the initial contextual accessibility and the length of the fit, so route share is not isolated as a cause; the theory's third candidate remedy, blocking the neighbour-to-encoder route, does not exist in this implementation.

C.8 Spectral Decoupling (K5)

Table 12: K5: Spectral Decoupling ($M = 32$, seeds 0–2): the evaluable strengths at L^* beside the $\kappa = 1$ run they modify, the verdicts and the censored strength.

λ	probe	probe − (κ=1)	reversed	recovery	recovery − (κ=1)
0 ($\kappa = 1$)	0.662±0.029	–	0.121±0.008	0.583±0.040	–
0.001	0.662±0.030	0.000	0.120±0.008	0.583±0.040	0.000
0.01	0.665±0.030	+0.003	0.114±0.008	0.586±0.042	+0.003

P52 (probe at least 0.05 above $\kappa = 1$ in every seed): largest paired difference $+0.003$ — not met
P53 (reversal accuracy above 0.50 in every seed): largest value 0.127; no strength's three-seed mean exceeds 0.50 either — not met
P54 (recovery at least 0.05 above $\kappa = 1$ in every seed): largest paired difference $+0.004$ — not met

$\lambda = 0.1$ censored: 0/3 seeds reached L^* within the $40{,}000$-step ceiling; no verdict, no recovery row.
Budget end (step 40,000, 3 seeds): cross-entropy 0.351±0.006, probe 0.911±0.016, reversed 0.298±0.066; not evidence about the penalty's effect.

Mean±sd over seeds 0–2; row $\lambda = 0$ is the registered $\kappa = 1$ comparator (not re-run). Differences are paired within seed against $\kappa = 1$ (columns: mean; verdict rows: the extreme seed), over the 2 evaluable strengths. Both gates passed before any outcome was read: without the penalty the modified code reproduces the registered run exactly, and each penalized run starts from the $\kappa = 1$ run's initial state bit for bit.

Registered before its runs (§27.5). Every intervention above changes how fast a part of the model learns; Spectral Decoupling (Pezeshki et al., 2021), the canonical remedy for gradient starvation, instead adds a penalty $\lambda \overline{F^2}$ on the logits to the registered $\kappa = 1$ run, nothing else changed, with $\lambda \in \{0.001, 0.01, 0.1\}$ fixed before the runs against the logit scale at matched fit (spanning 0.4% to 37% of the training loss there) and L^* read on the cross-entropy term alone. $\lambda = 0.1$ is censored and enters no verdict; at the two evaluable strengths *P52, P53 and P54 are not met* (Table 12). These verdicts concern this two-route synthetic construction, not the method: they do not refute Pezeshki et al. (2021), whose setting, architecture and failure mode differ, and no strength, seed or threshold was changed after the outcome.

D Experiment 3 details

D.1 Data, readiness, construction and arms

Material no longer printed here is archived verbatim in `docs/appendix_archive/D_exp3_*.tex`; registered analyses, thresholds, populations, outcomes and censoring are unchanged.

Data. QCL infrared images of a commercial breast-cancer tissue microarray (BR20832, US Biomax Inc.; 169 cores, one per donor; de-identified samples collected under donor consent and institutional approval at the originating institutions; used previously by Tang et al. (2021)) at $4\ \mathrm{cm}^{-1}$

and $4.25\,\mu$m pixels, 314 wavenumbers over 1016–1784 cm^{-1}, four tissue classes. We use fold 0 of the five-fold patient-level split used throughout (115 training, 28 validation, 26 test cores; the three sides disjoint in patients); preprocessing, principal directions, calibration and the readiness scan use training patients only. A pixel is the 942-dimensional vector of absorbance and its first and second derivatives after the companion pipeline's preprocessing, which includes per-core PCA denoising with 23 components before the derivatives are taken (numerical rank 23 per core, not in the pooled data, whose retained subspaces differ between cores). With a fixed seed, up to 600 (training) or 300 (validation, test) pixels per class per core give 75,297 training pixels from 79 cores, 11,024 validation from 20 and 8,038 test from 16; the pair is Cancer Epithelium ($+1$; 29,603 training pixels) against Cancer-Associated Stroma (-1; 21,923).

Readiness. The scan behind the two regimes of Section 5 covered four class pairs, bottleneck $K \in \{2, 4, 12, 32\}$ and both preprocessings, with random encoders fitted on one half of the training cores and evaluated on the other: after per-feature standardization a random encoder already reads the epithelium-versus-stroma contrasts, which lie along the top principal directions, while whitening (PCA, relative ridge 10^{-3}) removes that accessibility and leaves the oracle high (Table 13, top: the pair used, at $K = 12$).

Constructed context, model and arms. Beyond Section 5: the eight donors are drawn uniformly from the same split side without conditioning on the centre label, resampled once if a donor's core equals the centre's (so the donor distribution depends weakly on the centre's core); the cue $\gamma_d\,(c\,s + \tau\eta_d)\,e_d$ lies along the dth standardized principal direction $e_d = v_d$ with $\gamma_d = 3\sqrt{\lambda_d}$ (high accessibility) or the dth whitened coordinate with $\gamma_d = 30$, or 10 (an amendment prompted by the $\gamma = 30$ outcome and committed before its runs, §5) (3 and 1 in Appendix D.6); $s = +1$ is informative, $s = -1$ reversed, an independent random label context-random, and *spectral-only* drops the cue and keeps the donors. The four conditions of one seed share centres, donors and η; τ is calibrated per regime on training-core halves (Table 13). Training uses 24,000 class-balanced centres, the probe 6,000 disjoint training pixels, evaluation up to 6,000 validation and 6,000 test centres with donors from their own side. The encoder $W \in \mathbb{R}^{12\times 942}$ acts on the nine spectra and a ReLU head of width M on the 108 encoded coordinates, with a linear readout, trained as in Experiment 2 (margin loss, full batch, rate 10^{-3}, at most 40,000 steps, its thresholds and snapshot rule). Arms per regime (at $\gamma = 10$, `sp` and `ctxfree` at $M = 32$ only): `sp` at $M \in \{8, 32, 128, 512\}$; `headlr` at $\kappa \in \{1, 1/16, 1/256\}$, $M = 32$; `ctxfree` (context-random training) at $M \in \{8, 128\}$ (32 at $\gamma = 10$); `frozen` at $M = 32$; seeds 0–2. Recovery (Appendix C.2) starts from the random, $\kappa = 1$ and $\kappa = 1/256$ encoders at L^*.

D.2 Results

Table 13 gives every width and head rate run, per regime, as means over all five seeds (three registered, two added; its note marks the three-seed cells), and the retraining outcomes. Seeds 3 and 4 were added after the registered runs (registration section 11, the authors' record, first committed while those runs were in progress; Appendix H) and enter the table; the registered statements of Section 5 keep their three-seed values, and the added seeds agree on every contrast (rate responses $+0.032$ and $+0.054$ at $\gamma = 30$, $+0.086$ and $+0.082$ at $\gamma = 10$).

D.3 Five patient folds (extension, registration §12)

Under registration section 12 (the authors' record; Experiment 3's fold 1 ran before the section was first committed, the other seven fold blocks after it, Appendix H) the fold-0 protocols were run on the other four patient-level folds (Experiment 3: Table 14, top; Experiment 4: Table 20); registered verdicts are not recomputed. At low accessibility every tabulated contrast keeps its registered direction in $5/5$ folds (the $\gamma = 10$ recovery pair in all 15 fold–seed comparisons); in the high-accessibility control the paired differences change sign between folds with magnitudes at most 0.009. A later registered extension (§27.8) applied the blocking and centre-pretraining predictions within each of folds 1–4 (Table 14, bottom); every other number in Appendices D.4 and D.5 is fold 0's. Its first two launches (K8, K8b) were void under their own gate rules before any fold-1–4 run began (an identity-gate exclusion list that predated two wall-clock columns, with three unrelated tracked files uncommitted; then a commit check that compared the repository head with the code commit and

Table 13: Experiment 3, fold 0, validation patients, at $L^* = 0.30$. *Top*, readiness (three random encoders, $K = 12$, training-core halves) and calibration of the pair used. *Middle*, every width and head rate run, per regime, as means over all five seeds (three registered, two added; exceptions marked in the note). *Bottom*, recovery at $M = 32$ (Appendix C.2) from the random initial (`init`), fast (`kappa1`) and slow (`kappa256`) encoders.

regime	scan oracle	random-encoder probe	τ	centre oracle (fit/val)	context oracle
high accessibility (standardized)	0.969	0.955	1.475	0.969 / 0.959	0.969
low accessibility (whitened)	0.944	0.539	1.735	0.948 / 0.884	0.948

		high accessibility			low accessibility, $\gamma = 30$			low accessibility, $\gamma = 10$		
arm	M, κ	steps	gain	rev.	steps	gain	rev.	steps	gain	rev.
`sp`	8	184	$+0.002$	0.094	220	$+0.026$	0.077	–	–	–
`sp`	32	144	-0.002	0.086	128	$+0.036$	0.070	642	$+0.088$	0.067
`sp`	128	112	$+0.003$	0.099	105	$+0.038$	0.071	–	–	–
`sp`	512	76	$+0.004$	0.097	73	$+0.008$	0.066	–	–	–
`headlr`	32, 1/16	236	-0.002	0.084	209	$+0.059$	0.072	1,555	$+0.143$	0.066
`headlr`	32, 1/256	249	-0.001	0.083	220	$+0.061$	0.072	1,928	$+0.157$	0.067
`ctxfree`	8	1,226	$+0.016$	0.923	17,710	$+0.372$	0.924	–	–	–
`ctxfree`	128	878	$+0.014$	0.925	15,457	$+0.358$	0.908	–	–	–
`ctxfree`	32*	–	–	–	–	–	–	16,003	$+0.350$	0.916
`frozen`	32	498	0.000	0.092	451	0.000	0.073	2,224	0.000	0.081

	high accessibility			low accessibility, $\gamma = 30$			low accessibility, $\gamma = 10$		
encoder	probe	recovered	[min, max]	probe	recovered	[min, max]	probe	recovered	[min, max]
`init`	0.925	0.916	[0.888, 0.953]	0.562	0.502	[0.445, 0.546]	0.562	0.502	[0.431, 0.556]
`kappa1`	0.923	0.931	[0.909, 0.959]	0.598	0.519	[0.447, 0.571]	0.651	0.561	[0.477, 0.636]
`kappa256`	0.924	0.936	[0.926, 0.959]	0.623	0.531	[0.458, 0.593]	0.719	0.659	[0.550, 0.755]

$n = 5$ seeds per cell (three registered, two added; $^*n = 3$, seeds 0–2); every run reached L^* ($5/5$ per cell, $3/3$ for *), so no cell is censored (the censored blocked slow arm is in Appendix D.4). gain: probe at L^* minus the initial probe (validation centres); rev.: reversal accuracy. `headlr` $\kappa = 1$ is the `sp` $M = 32$ run. –: not run at seeds 0–2; at $\gamma = 10$, `sp` $M \in \{8, 128, 512\}$ and `ctxfree` $M \in \{8, 128\}$ were run at the added seeds 3–4 only ($n = 2$ each, all reached L^*; repository only). Over the printed cells: sd at most 0.060 (gain) and 0.024 (rev.); context-random accuracy 0.497–0.509 (informative context, `frozen`) and 0.905–0.925 (`ctxfree`); spectral-only 0.491–0.559 (informative context, `frozen`); test-patient reversal (used for no choice) within 0.028 of validation. Bottom: $n = 5$, all retraining runs completed; probe: the encoder's probe (for `init`, the initial probe); recovered: mean reversal accuracy of the retrained head and its per-seed range over the five seeds (Section 5 quotes the registered seeds 0–2; sd at most 0.078; test within 0.035); properties of this procedure, checkpoint and budget, not ceilings. The class-mean contrast has cosine 0.50 with the first standardized principal direction. Registered seeds 0–2 (the values Section 5 quotes): high accessibility, informative reversal 0.08–0.10, comparator 0.93, frozen 0.09; $\gamma = 30$, informative gain at most 0.046 against the comparator's 0.34–0.35, reversal 0.07–0.08 against 0.91–0.92, frozen 0.08, rate response (probe gain, $\kappa = 1/256$ minus 1, $M = 32$) $+0.013$; $\gamma = 10$ rate response $+0.058$; recovery 0.90–0.96 (high), 0.45–0.56 ($\gamma = 30$), 0.55–0.70 against 0.48–0.60 ($\gamma = 10$, slow against fast).

could not pass), the fold-0 reproduction otherwise matching exactly in both; the third launch, K8c, changed only those gate mechanics and the output directories, with predictions, populations and thresholds unchanged, and is the one reported (pre-registration file, §§27.8, 27.8b, 27.8c; the launch records are in the release).

D.4 Blocking the neighbour-to-encoder route (registered extensions C8, C8b)

Registered before their runs (pre-registration file, §21.2; C8b in §22, registered after the C8 censoring below was observed): in the setting of Table 15, two arms ($\kappa = 1$, $1/256$) detach the eight neighbours' encoder outputs before the head as in Section 5, which at fixed parameters leaves logits and head-parameter gradients unchanged (an identity test on a random batch, in single and double precision, is part of the registration); initialization and data are those of the ordinary arms. The slow blocked arm did not reach L^* within the registered $40{,}000$-step ceiling in any seed (final losses 0.31, 0.41, 0.36), which left the registered comparison not evaluable; an extension (C8b), registered after this censoring had been recorded and before its runs, raised its ceiling to $400{,}000$ steps with the predictions unchanged, and the rerun reproduces the first run at every shared snapshot. *P16: met* — at matched fit the blocked slow arm exceeds the ordinary slow arm by at least 0.10 in mean probe gain and in mean fresh-head recovery, with a positive difference in every seed. *P17: met* — the interaction between blocking and slowing is positive in mean and in at least two of three seeds for both quantities. At the fast head rate blocking does not consistently improve either quantity (probe -0.03, recovery -0.01 against the ordinary fast arm). The arms are matched on training loss, not on steps: the blocked slow arm takes $43{,}406$–$65{,}178$ steps against 140–226, so its encoder also receives far more centre-derived updates; this comparison does not separate the gradient route from the step count. A fixed-step control (K3), registered after these outcomes were read (pre-registration

Table 14: Five patient partitions of one cohort (fold 0 registered; seeds 0–2 per fold; validation patients; $L^* = 0.30$; pairs formed within (fold, seed) at $M = 32$). *Top*, unblocked Experiment 3 contrasts. *Bottom*, K8c. The partitions share training cores and are not independent cohorts.

Experiment 3, unblocked arms	mean	[min, max] over folds	folds	pairs
Low accessibility (whitened), $\gamma = 30$				
probe gain, informative context (sp, $M \in \{8, 32, 128, 512\}$; $n = 12$)	+0.022	[+0.015, +0.037]	5/5	
probe gain, uninformative comparator (ctxfree, $M \in \{8, 128\}$; $n = 6$)	+0.360	[+0.327, +0.377]	5/5	
reversal accuracy, informative context ($n = 12$)	0.103	[0.074, 0.126]		
reversal accuracy, uninformative comparator ($n = 6$)	0.918	[0.907, 0.933]		
paired Δ probe gain, $\kappa = 1/256$ minus $\kappa = 1$ ($M = 32$)	+0.018	[+0.006, +0.031]	5/5	14/15
paired Δ recovered reversal, $\kappa = 1/256$ minus $\kappa = 1$	+0.008	[+0.003, +0.012]	5/5	13/15
Low accessibility, $\gamma = 10$ *(*$M = 32$ *only: the support common to all five folds)*				
paired Δ probe gain, $\kappa = 1/256$ minus $\kappa = 1$	+0.083	[+0.058, +0.113]	5/5	15/15
paired Δ recovered reversal, $\kappa = 1/256$ minus $\kappa = 1$	+0.097	[+0.077, +0.141]	5/5	15/15
High accessibility (standardized): the control regime				
paired Δ probe gain, $\kappa = 1/256$ minus $\kappa = 1$	0.000	[-0.009, +0.003]	4/5	11/15
paired Δ recovered reversal, $\kappa = 1/256$ minus $\kappa = 1$	0.000	[-0.001, +0.004]	3/5	8/15

mean: equal-weight mean over the five folds; [min, max]: the per-fold extremes; folds: folds with the stated sign; pairs: (fold, seed) pairs with a positive difference. No cell is censored (all 438 registered runs at these seeds reached L^*; all 135 recovery runs completed); no $\gamma = 10$ ctxfree width is common to all folds. Recovered values are properties of this retraining procedure and budget, not ceilings; per-fold values are archived.

K8c, fold forms of the fold-0 clauses	fold means [min, max]	15 fold–seed values [min, max]	folds†
Blocking the neighbour-to-encoder route (C8/C8b), at $\kappa = 1/256$			
P61(i) = P16, probe gain, blocked − open	$[+0.268, +0.318]$	$[+0.168, +0.365]$	5 / 0 / 0
P61(ii) = P16, recovery, blocked − open	$[+0.343, +0.417]$	$[+0.307, +0.442]$	5 / 0 / 0
P62 = P17, interaction, probe	$[+0.314, +0.345]$	$[+0.262, +0.389]$	5 / 0 / 0
P62 = P17, interaction, recovery	$[+0.361, +0.438]$	$[+0.323, +0.451]$	5 / 0 / 0
steps to L^*, blocked slow	–	[43,406, 114,321]	
steps to L^*, open slow	–	[140, 1,681]	
Centre pretraining with a fresh or a prepared head (C5), at $\kappa = 1$			
P63 = P20, probe, $L^* - W_{\mathrm{pre}}$	–	$[-0.004, +0.002]$	5 / 0 / 0
P64 = P21, fresh-head reversal	–	[0.106, 0.138]	5 / 0 / 0
P65 = P22, reversal, prepared − fresh	$[+0.467, +0.549]$	$[+0.425, +0.583]$	5 / 0 / 0
S1, prepared-head reversal (secondary)	–	[0.563, 0.689]	5 / 0 / 0
P66(a) = P23(a), recovery, fresh − ordinary	–	$[+0.335, +0.470]$	5 / 0 / 0
P66(b) = P23(b), recovery, fresh − W_{pre}	–	$[-0.003, +0.002]$	5 / 0 / 0

$\gamma = 30$ low-accessibility construction, $M = 32$. P61–P66 apply the fold-0 clauses of P16–P17 (Appendix D.4) and P20–P23 with the separately counted check S1 (Appendix D.5) within each fold, paired within seed; the ranges (from unrounded values; fold means only where the clause has a mean condition) decide every per-fold verdict. †Folds met / not met / not evaluable; a prediction holds in the five-fold sense when it is met in at least four folds; no cell in any fold is censored, void or degenerate. Fold 0 is read from the recorded runs, its verdicts not re-scored; folds 1–4 were run after an identity gate reproduced every fold-0 object exactly, against each fold's recorded ordinary arms and recovery rows. P63 and P66(b) are tolerance clauses met on near-zero quantities: the quantity does not move; they record no effect. In every fold the blocked slow arm needs far more updates than the open slow arm to reach the matched loss, so these are matched-fit, not matched-step, comparisons.

file, §27.3), re-ran the four arms with the encoder also saved at shared training steps (Table 15, bottom). At the $4{,}000$-step checkpoint the recovery advantage recorded at matched fit is not present (the blocked slow arm is behind the open one in every seed, -0.086 on average); at $20{,}000$ steps it is present, $+0.121$ on average and positive in every seed. The primary probe and recovery predictions were not met; at $20{,}000$ steps the secondary recovery prediction was met, while the positive probe difference missed the registered 0.10 mean-effect threshold. The reading is the registered one: context can continue adapting through the shared encoder when only the head is slowed, and blocking that route restores the rate response in this constructed setting and in ≥ 4 of the five patient partitions (Table 14). It is an intervention in one setting, not a rule; the blocked slow classifiers remain context-reliant at matched fit (reversal 0.14); and, unlike the toy's separate contextual coefficient, the experimental encoder stays shared, so blocking neighbour-derived gradients does not freeze neighbour representations against changes induced by centre-derived updates. Corollary A.1 is the exact toy analogue.

D.5 Centre pretraining with two head preparations (registered extension C5)

Registered before its runs (pre-registration file, §21.5, amendments §21A in force), in the setting of Table 16. The encoder is first trained on the training centres alone to its own loss threshold 0.30, its readout discarded, giving W_{pre}; joint training on the ordinary patches then starts from a fresh random head or from a head prepared alone, with W_{pre} frozen, on the decorrelated-context twin of the same patches. The two arms share an initialization (identity check: maximum absolute

Table 15: Blocking the neighbour-to-encoder gradient route: $\gamma = 30$ low-accessibility tissue construction, $M = 32$, fold 0, validation patients, seeds 0–2. *Top*, C8 and C8b at matched fit, with the per-seed P16 and P17 differences. *Bottom*, K3, the fixed-step control: per-seed and mean differences D, I and R at 4,000 and 20,000 training steps with their clause verdicts, and the four arms' mean training losses at those steps.

neighbour route	κ	probe init.	probe L^*	gain	steps to L^*	reversal L^*	recovery
open	1	0.582	0.614	$+0.032$	95/137/107	0.072	0.519
open	1/256	0.582	0.628	$+0.046$	140/226/188	0.075	0.526
blocked	1	0.582	0.581	0.000	359/500/324	0.076	0.506
blocked	1/256	0.582	0.909	$+0.327$	43,406/65,178/56,922	0.142	0.909
P16 differences per seed: gain +0.285, +0.233, +0.325; recovery +0.399, +0.384, +0.365							
P17 interactions per seed: probe +0.279, +0.318, +0.344; recovery +0.398, +0.399, +0.389							

Means over seeds (steps per seed), read at the first step with training loss below $L^* = 0.30$; recovery: reversal accuracy after the protocol of Appendix C.2 (random encoder: 0.505).

quantity	step	seed 0	seed 1	seed 2	mean	registered clause: verdict
D	4,000	-0.027	-0.150	-0.125	-0.101	P45 (i), (ii), (iii) not met
D	20,000	$+0.097$	$+0.032$	$+0.060$	$+0.063$	P45b (i) met, (ii) not met
I	4,000	-0.002	-0.005	-0.050	-0.019	P48 (i) not met
I	20,000	$+0.047$	$+0.080$	$+0.074$	$+0.067$	P48 (ii) met
R	4,000	-0.059	-0.102	-0.097	-0.086	P46 (i), (ii) not met
R	20,000	$+0.150$	$+0.105$	$+0.109$	$+0.121$	P46b (i), (ii) met
D, R	55,000	not defined: the open arms stop at 40,000				P47 not evaluable $(0/3)$
mean loss	4,000	open fast 0.131; open slow 0.144; blocked fast 0.140; blocked slow 0.682				
mean loss	20,000	open fast 0.128; open slow 0.133; blocked fast 0.129; blocked slow 0.499				

An identity gate required every value the re-runs share with the recorded runs to be equal; it passed for all four arms. D is the probe of the blocked slow arm minus that of the open slow arm; $I = D - D_{\text{fast}}$, with D_{fast} the same difference between the blocked and open fast arms (so $D_{\text{fast}} = D - I$); R is the blocked-slow minus open-slow difference in recovery, the protocol of Appendix C.2 applied to the checkpoint saved at the step (random encoder: 0.445 / 0.546 / 0.523); all are formed within seed from unrounded values. Registered clauses, fixed before the run, no tolerance: at 4,000 steps P45 (i) $D > 0$ in every seed, (ii) mean $D \geq 0.10$, (iii) $D \geq 0.10$ in every seed; at 20,000 P45b (i) and (ii), the same; P46 (i) $R > 0$ in every seed and (ii) $R \geq 0.10$ in every seed and in the mean at 4,000, P46b the same at 20,000; P48 (i) and (ii), $I > 0$ in the mean and in at least two of three seeds at 4,000 and 20,000. P47, the comparison at 55,000 steps, is not evaluable because only the blocked slow arm, whose ceiling is 400,000 steps, reaches that step. The step-40,000 values were known before this control was registered and are reproduced exactly (means $D = +0.130$, $I = +0.052$, against the matched-fit probe interaction $+0.314$ of Appendix D.4); the matched-fit verdicts of Appendix D.4 are not revised. Presentation amendment (24 September): the registered display listed every shared step; steps 0, 200, 500 and 55,000, the D_{fast} rows and the per-arm values are archived; no analysis or verdict changed.

Table 16: C5: centre pretraining with a fresh and with a prepared head ($\gamma = 30$ low-accessibility tissue construction, $M = 32$, fold 0, validation patients, seeds 0–2): means over seeds (steps per seed) and the per-seed differences behind P20–P23.

encoder, head start ($\kappa = 1$)	probe start	probe L^*	steps to L^*	reversal L^*	recovery
ordinary: random W_0, fresh head	0.582	0.614	95/137/107	0.072	0.519
W_{pre}, fresh head	0.915	0.915	64/68/96	0.116	0.913
W_{pre}, prepared head	0.915	0.915	3/4/4	0.665	0.914
Per seed: P20 probe difference $+0.002$, 0.000, -0.002; P21 reversal 0.106, 0.130, 0.114					
P22 prepared $-$ fresh $+0.583$, $+0.513$, $+0.549$; P23 fresh $-$ $\kappa = 1$ $+0.470$, $+0.341$, $+0.374$					
P23 fresh $-$ pretrained -0.001, $+0.001$, -0.002					

Pretraining to loss 0.30 took 4490/3659/3840 steps and head preparation to loss 0.40 7334/6400/6749. Rows are read at the first step with training loss below $L^* = 0.30$; "probe start" is the LDA centre probe at the start of joint training; recovery is the reversal accuracy after the protocol of Appendix C.2; from W_{pre} without joint training it is 0.914 and from the random encoder 0.505. The registered decorrelated-context reference (`ctxfree`, $M \in \{8, 128\}$, same seeds; no recovery row is registered for it) reaches reversal 0.876–0.940 per seed at L^* (Table 13 gives its five-seed means).

difference 0). All six joint runs reached L^* with positive joint-training duration, so the unconditional and positive-duration populations coincide at $3/3$; all nine retraining runs completed; nothing is censored. Pretraining raises the mean centre probe from 0.582 to 0.915. *P20: met* (probe at $L^* \geq$ probe of $W_{\text{pre}} - 0.05$, every seed): joint training leaves it there. *P21: met* (reversal at $L^* \leq 0.50$, every seed), the informative result: with a fresh head the fitted classifier is still context-reliant, so an accessible centre cue is not sufficient. *P22: met* (prepared minus fresh reversal ≥ 0.10 in the mean and > 0 in every seed): preparing the head on decorrelated context raises reversal at L^* to 0.643–0.689, also meeting the separately registered check that the prepared reversal is ≥ 0.50 in every seed; we report it as the intervention that changed the original classifier, in this construction and in ≥ 4 of the five patient partitions (Table 14), a linear-probing-then-fine-tuning recommendation with closely related prior work (Kumar et al., 2022; Ye et al., 2023), not a novelty claim. *P23: met in both clauses* (each seed: recovery $\geq$ that of the ordinary $\kappa = 1$ encoder $+0.10$, and $\geq$ that of W_{pre} -0.05): a fresh head retrained on decorrelated context reaches 0.913 on average from the

jointly trained encoder, within 0.002 of W_{pre} itself, against 0.52 from the ordinary and 0.50 from the random encoder. Pretraining therefore preserves recoverability without curing reliance, in this construction and in ≥ 4 of the five patient partitions. The preparation's advantage is a matched-fit quantity that does not persist: by the end of training (loss 0.10) the prepared arm's reversal accuracy has fallen from 0.643–0.689 to 0.139–0.146, against 0.067–0.071 for the fresh arm, and even at L^* it remains reliant (context-randomized 0.786–0.816, in distribution 0.932–0.944; the decorrelated-context reference is in the note to Table 16). This is one construction and one class pair and, beyond the four registered comparisons, one fold; it is not a universal cure, and not an instance of the theorem's head-start lever: h_0 is not the toy's spectral coefficient a_0, and nothing here establishes $a_0 \geq m/2$.

D.6 Amplitude sweep (registered predictions C12, with contingency)

Table 17: C12: the four-level contextual-amplitude sweep on constructed tissue. Verdicts: P38 met; P36, P37 and the toy-calibrated P41 not met.

γ	$\bar{\rho}$	R per seed	mean R	G per seed	mean G	comparator per seed	`init`
$30^{\dagger}$	0.9887	$+0.002, +0.023, +0.016$	$+0.0134$	$+0.011, +0.002, +0.008$	$+0.0070$	–	0.505
$10^{\dagger}$	0.9671	$+0.030, +0.049, +0.095$	$+0.0577$	$+0.073, +0.097, +0.066$	$+0.0789$	0.352, 0.337, 0.360	0.506
3	0.8984	$+0.130, +0.116, +0.137$	$+0.1277$	$+0.176, +0.157, +0.189$	$+0.1742$	0.354, 0.341, 0.359	0.519
$1^{\ddagger}$	0.7489	$+0.163, +0.049, +0.109$	$+0.1068$	$+0.170, +0.070, +0.109$	$+0.1163$	0.354, 0.341, 0.362	0.530

Constructed whitened tissue neighbourhoods at four cue amplitudes γ, $M = 32$, fold 0, CancerEpi vs CAS, validation patients, seeds 0–2 in seed order. $\bar{\rho}$ is the three-seed mean initialisation route share, measured before any run. R (rate response) is the probe gain at $\kappa = 1/256$ minus the probe gain at $\kappa = 1$, both read at the first training-loss crossing of $L^* = 0.30$; G (recovery gain) is the reversal accuracy after retraining a fresh head on decorrelated context, $\kappa = 1/256$ encoder minus $\kappa = 1$ encoder; the comparator is the probe gain of the decorrelated-context arm at $\kappa = 1$; `init` is the random-encoder recovery baseline. † pre-existing registered levels, reproduced from the registered files and not re-run. ‡ the $\gamma = 1$ `headlr` runs are the registered contingency (text): the extended runs reproduce the censored ones exactly at every shared snapshot and every shared trajectory row up to step $40{,}000$ (max |difference| $= 0$). Its `ctxfree` comparator was not re-run. Amplitude also changes the absolute readability of the cue and the length of the fit, so these columns do not isolate the route share. Steps to L^* ($\kappa = 1$ | $\kappa = 1/256$, seeds 0/1/2): $\gamma = 30$: 95/137/107 | 140/226/188; $\gamma = 10$: 528/666/581 | 1,217/1,809/1,712; $\gamma = 3$: 2,374/2,744/2,449 | 12,612/16,889/16,741; $\gamma = 1$: 7,320/7,929/7,061 | 74,869/84,386/80,509 — the fit-length confound reported beside every C12 number.

Registered before its runs (pre-registration file, §24, amendments §24A; toy-calibrated predictions §24.3 recorded before any outcome; contingency §24.1-1): the $\gamma = 30$ and $\gamma = 10$ constructions are extended downward to $\gamma = 3$ and $\gamma = 1$ at $M = 32$, fold 0, seeds 0–2, the design's gates re-run and passed; the predictor fixed before the runs is the initialization neighbour-route share $\bar{\rho}$. All three $\kappa = 1/256$ runs at $\gamma = 1$ were censored at the $40{,}000$-step ceiling; the registration's own contingency re-ran that level with a $400{,}000$-step ceiling and reproduced the censored runs exactly (Table 17, ‡); nothing is censored afterwards. The full clause text is in the registration and the archive; the deciding criteria and outcomes are these. *P38*, the control (at $\gamma = 3$ and 1 the decorrelated-context comparator reaches probe gain ≥ 0.30 at L^* in every seed; a failing level would not be interpreted): *met*, so no level is gated out. *P36* on the rate response R and *P37* on the recovery gain G (three-seed means) each required (a) no fall of more than 0.01 between consecutive levels as $\bar{\rho}$ decreases, (b) the $\gamma = 1$ value at least twice the recorded $\gamma = 10$ value ($R \geq 0.115333$, $G \geq 0.157778$), and (c) the $\gamma = 1$ value above the $\gamma = 10$ value in at least two of three seeds: *not met* — (a) holds over the first three levels and fails at the fourth (consecutive differences $+0.044$, $+0.070$, -0.021 for R and $+0.072$, $+0.095$, -0.058 for G, the last outside the tie band), (b) fails ($R(\gamma = 1) = +0.1068$, $G(\gamma = 1) = +0.1163$), and (c) holds in two of three seeds for both quantities, one of them by 0.000167. *P41*, calibrated from the toy on the two existing levels (fitted anchors 0.013444 at $\gamma = 30$ and 0.057667 at $\gamma = 10$, not predictions; predicted $R = 0.032303$ at $\gamma = 3$ and 0.004876 at $\gamma = 1$), required the ordering of the four means (the $\gamma = 30$ versus $\gamma = 1$ pair declared unresolved), the maximum at $\gamma = 10$, and $0.013444 < R(\gamma = 3) < 0.057667$ with $R(\gamma = 1) < R(\gamma = 3) - 0.01$: *not met* on all three clauses — the $(\gamma = 10, \gamma = 3)$ pair inverts by 0.070 against a predicted $+0.025$, the argmax is $\gamma = 3$, and $R(\gamma = 3)$ is not below 0.057667.

As in Appendix C.7, neither the monotone ordering by initialization-time route share nor the toy-calibrated peak was supported. Both quantities rise and then fall, their maximum being the interior level $\gamma = 3$: a maximum among four measured points, not an established turning point. From $\gamma = 3$ to $\gamma = 1$ the difference shrank because the *ordinary* arm improved more, not because slow training produced a worse representation — the slow arm's values at $\gamma = 1$ are the best it attains in the sweep (probe gain $+0.346$, recovered reversal 0.933 against a random-encoder baseline of 0.530) while the ordinary arm gained $+0.097$ in probe and $+0.142$ in recovery against the slow arm's $+0.076$ and

$+0.084$ — which changes no verdict and is not a normalized rescue; C12 registered no normalized clause. Amplitude is confounded with cue readability and fit length (note to Table 17).

D.7 Initialization route share across settings (exploratory C8c)

The share of the encoder gradient carried by the neighbour tokens at initialization, $\|g_{\mathrm{n}}\|/(\|g_{\mathrm{c}}\| + \|g_{\mathrm{n}}\|)$, was measured once on the first (full) training batch of each setting, at the registered initialization, $M = 32$, seeds 0–2, with the route decomposition of Appendix D.4 (the two route gradients sum to the ordinary encoder gradient to 5.4×10^{-7}). It is exploratory and retrospective: which settings had met the registered predictions was already known when the statistic was fixed. This registered retrospective ordering (shares higher in every seed in the failed settings, tissue and Pavia at $\gamma = 30$, than in the held ones, Experiment 2 linear at $M = 32$ and tissue at $\gamma = 10$) fails under its governing population, which also includes Pavia $\gamma = 10$ under the registration's wording: Pavia $\gamma = 10$ failed the rate criterion despite route shares 0.9586–0.9663, overlapping the successful tissue $\gamma = 10$ range 0.9583–0.9719. The full setting-by-seed record is archived; this statistic supplies no cross-setting decision rule; no rescue statistic was sought. The overlap is a witness against the ordering, not a significance test.

E Experiment 4 details: natural context

E.1 Data, arms and registered rules

Experiment 4 is a usage test, not a mechanism test: no cue is constructed or calibrated. Cores, patient-level fold-0 split and four classes are those of Appendix D. A patch is a labelled centre pixel with its eight recorded neighbours, all nine inside the tissue mask; up to 150 (training) or 100 (validation, test) patches per core and class, sampled with a fixed seed, give $19{,}650$ training patches from 79 cores, $3{,}700$ validation from 20 and $2{,}700$ test from 16. No labelled neighbour disagrees with its centre (88% share its label, 12% are unlabelled), a property of the sample that does not imply that neighbouring measurements carry no additional class information or cannot compete during optimization. The non-denoised copy of the spectra (full rank; the registered primary set) and the companion pipeline's PCA-23-denoised copy, a replication (closer to the preprocessing of O'Leary et al. (2026); identical arms and predictions, registration section 8.4) are analysed separately and never pooled. Predictions, arms and disposition rules are those of registration section 8, recorded by the authors before the first run and first committed together with the experiment's code and first results (Appendix H); its two amendments (the class-weighted loss; the no-bottleneck arms, added after the authors' code was checked) are recorded there.

Stage A fits a shrinkage Fisher discriminant, balanced logistic regression and a 256-unit ReLU MLP (three seeds) to standardized training centres on k principal components (PCA on training patients), scored on validation and test patients without selection; the registered gate G1 requires a macro-F1 gain of at least 0.02 from $k = 16$ to the full spectrum on *both* sides. *Stage B* trains the Experiment 3 model with four logits and class-weighted cross-entropy (linear encoder $942 \to 12$, or $16 \to 12$ in the 16-PC arms, on all nine pixels; ReLU head, $M = 32$; full-batch gradient descent at rate 10^{-3}; paired initialization within seed; three seeds) until $40{,}000$ steps or the declared stopping loss 0.10 (budget end). The comparator, *shuffled*, replaces the neighbours by random training centres drawn without conditioning on the centre label; *frozen* keeps a random encoder; $\kappa = 1/256$ multiplies the whole head's rate; *no bottleneck* (amendment 8.5) is the identity encoder, the analogue of the "fixed" models of O'Leary et al. (2026). Each centre is evaluated with its natural (*iid*), *context-random* (centres from other cores, independent labels) or *homogeneous* (the centre copied) neighbours. The probe, a four-class discriminant on the encoder output of centre spectra fitted on $4{,}000$ held-out training centres, is reported as its gain over the arm's initialization; retraining follows Appendix C.2; Stage B verdicts use validation patients only. No one threshold in $\{0.30, 0.40, 0.50\}$ was reached by every arm, so the registered matched-fit comparison is unavailable; we report attainment with its denominators and the planned budget-end comparison, whose promotion to primary status occurred after the outcomes were known (registration 8.6). Rows and tables moved out of this appendix are in the release (`docs/appendix_archive/E_*.tex`).

Table 18: Experiment 4 on tissue, fold 0. (a) $O(\cdot)$: MLP, mean±sd over three seeds. (b) Mean±sd over n seeds; *loss*: training loss (0.10: stopped at the declared loss before 40,000 steps); *reached* $a/b/c$: seeds reaching $L = 0.5, 0.4, 0.3$; *retrained*: context-random macro-F1 of a fresh head (Appendix C.2) on the budget-end encoder or on the one a lower row names (–: no fresh-head recovery was run for that arm; none without an $L = 0.3$ checkpoint; tenfold ceiling: Table 19); its test column is reported and never used for a verdict. (c) $O(k)$: validation macro-F1 of the MLP (the best family on both copies) at k components; nat, nat16, shuf, shuf16: natural, natural 16-PC, shuffled, shuffled 16-PC; natfull, shuffull: no bottleneck. Archived: k-sweep, LDA and LR levels, MLP test levels, steps, absolute probe, Stage B test columns, $L = 0.5$ rows, per-seed values, retrained iid.

(a) Stage A gate G1: gain Δ from $k = 16$ to the full input (validation / test)

copy	LDA Δ	LR Δ	MLP Δ	MLP $O(16)$	MLP $O(942)$
non-denoised	+0.091 / +0.108	+0.078 / +0.087	+0.066 / +0.084	0.640±0.012	0.705±0.002
PCA-23-denoised	+0.121 / +0.264	+0.175 / +0.273	+0.139 / +0.249	0.665±0.005	0.804±0.002

(b) Stage B at budget end, validation patients, with fresh-head recovery

arm	loss	reached	probe gain	iid	ctx-random	retrained	retrained (test)	n
non-denoised copy								
natural, no bottleneck	0.10±0.00	3/3/3	0.000±0.000	0.744±0.005	0.249±0.007	–	–	3
natural	0.19±0.00	3/3/3	0.150±0.025	0.732±0.009	0.234±0.003	0.681±0.010	0.617±0.008	3
natural, $\kappa = 1/256$	0.48±0.02	3/0/0	0.106±0.024	0.652±0.012	0.296±0.018	0.644±0.008	0.604±0.002	3
natural, 16 PCs	0.42±0.00	3/0/0	0.031±0.020	0.657±0.004	0.201±0.024	0.594±0.012	0.568±0.002	3
shuffled, no bottleneck	0.11±0.00	3/3/3	0.000±0.000	0.675±0.005	0.661±0.006	–	–	3
shuffled (comparator)	0.34±0.00	3/3/0	0.148±0.012	0.652±0.008	0.680±0.006	0.688±0.007	0.631±0.007	3
shuffled, 16 PCs	0.53±0.00	0/0/0	0.015±0.014	0.541±0.004	0.583±0.011	–	–	3
natural, frozen encoder	0.63±0.03	0/0/0	0.000±0.000	0.576±0.015	0.264±0.022	–	–	3
natural, random-init encoder	–	–	–	–	–	0.494±0.036	0.481±0.044	3
natural, encoder at $L = 0.3$	–	–	–	–	–	0.678±0.006	0.624±0.005	3
PCA-23-denoised copy								
natural, no bottleneck	0.10±0.00	3/3/3	0.000±0.000	0.744±0.001	0.280±0.010	–	–	3
natural	0.10±0.00	3/3/3	0.142±0.002	0.727±0.004	0.274±0.018	0.706±0.003	0.791±0.005	3
natural, $\kappa = 1/256$	0.37±0.02	3/3/0	0.155±0.015	0.697±0.010	0.299±0.018	0.702±0.010	0.698±0.010	3
natural, 16 PCs	0.37±0.01	3/3/0	0.024±0.024	0.683±0.006	0.187±0.014	0.595±0.011	0.584±0.002	3
shuffled, no bottleneck	0.10±0.00	3/3/3	0.000±0.000	0.717±0.018	0.715±0.004	–	–	3
shuffled (comparator)	0.18±0.01	3/3/3	0.147±0.009	0.703±0.003	0.722±0.004	0.720±0.005	0.782±0.007	3
shuffled, 16 PCs	0.52±0.00	0/0/0	0.011±0.021	0.564±0.024	0.588±0.007	–	–	3
natural, frozen encoder	0.52±0.02	1/0/0	0.000±0.000	0.594±0.026	0.261±0.016	–	–	3
natural, random-init encoder	–	–	–	–	–	0.503±0.018	0.523±0.053	3
natural, encoder at $L = 0.3$	–	–	–	–	–	0.712±0.004	0.705±0.005	3
shuffled, encoder at $L = 0.3$	–	–	–	–	–	0.726±0.001	0.732±0.008	3

(c) Registered verdicts at budget end, validation patients

prediction	registered criterion	non-denoised	PCA-23
P7	$\lvert$iid(nat) − iid(nat16)$\rvert < 0.02$ (uninformative if both > 0.97)	not met	not met
P8, clause 1	ctx-rnd(nat) $\leq O(942) - 0.10$	met	met
P8, clause 2	ctx-rnd(shuf) $\geq O(942) - 0.03$	met	**failed**
P9, head-level	retrained ctx-rnd(nat encoder) $\geq$ ctx-rnd(shuf) -0.03	met	met
P9, encoder-level	retrained short by > 0.03 while ctx-rnd(shuf) $\geq O(942) - 0.03$	not met	not met
P10, numerical	iid(shuf) − iid(shuf16) $\geq 0.6\,[O(942) - O(16)]$ (G1 passed)	met	met
P10, asymmetry	the numerical clause together with P7	not established	not established
P11, clause 1	$\lvert$iid(nat) − iid(natfull)$\rvert < 0.02$	met	met
P11, clause 2	$\lvert$ctx-rnd(shuf) − ctx-rnd(shuffull)$\rvert < 0.03$	met	met
secondary (rate)	probe gain and retrained ctx-rnd at $\kappa = 1/256$ above $\kappa = 1$, every seed	not met	not met

E.2 Results on the two copies

G1 is met in all twelve comparisons (Table 18(a)), so sixteen principal components are not enough for per-pixel classification of these spectra by these classifiers. The dependence on k is not monotone (components 9–16 lower test macro-F1 for all three classifiers; the best non-denoised MLP value is at $k = 128$), so the gate concerns these classifiers at two operating points, not an information ordering in k; the top 16 components carry 81.7% of the standardized variance.

The copies are judged separately (Table 18(c)). *P7* fails on both. At budget end the full-spectrum model is also the better fitted (non-denoised training loss 0.19 against 0.42); at $L = 0.5$, the lowest loss both reach within the budget on the non-denoised copy, the gap is absent (0.643 against 0.636; on the denoised copy, at the same $L = 0.5$, 0.629 against 0.640; rows archived), and the registered runs do not isolate whether it reflects convergence or fitting speed. The budget-end comparator is also less fitted than the natural-context model it is set against in P8 and P9 (0.34 against 0.19); Section E.3 reads both at $L^* = 0.30$. For the values quoted in Table 1 and Section 6: at budget

end on the non-denoised copy the natural- and shuffled-context probes are 0.672 and 0.670 (validation macro-F1), and a fresh head retrained on decorrelated context from the natural-context encoder scores 0.673 in distribution against the joint model's 0.732. *P8*'s first clause holds on both copies; its second fails on the denoised copy, where the jointly trained comparator stays 0.08 below the per-pixel MLP, as does the no-bottleneck comparator: the gap persists without a bottleneck under this protocol, so the bottleneck alone does not explain it, and architecture and optimization effects are not separated here. As registered, P8 alone does not establish a shortcut (natural neighbours are almost always same-class, so part of the drop is legitimate noise averaging); the registration intended P9 and P10 to carry that claim. *P9* reads head-level on both copies, but the reading is checkpoint-specific and does not show that the two encoders learned the same features or that the encoder retained all task-relevant information. *P10*'s numerical clause holds on both copies, but its compression-asymmetry reading is not established, because the natural-context model loses from PCA-16 too (P7). What is supported is therefore comparator-level recovery under the stated retraining procedure (P9), alongside sensitivity to context-random replacement (P8). *The secondary rate prediction* fails on both copies; why slowing the head does not help is not identified here. *P11*, the comparison O'Leary et al. (2026) actually made (a learned linear bottleneck against none), reproduces on both copies within the registered validation tolerance, although on the denoised test patients the bottleneck model is 0.037 higher — a difference, not an equivalence — and the no-bottleneck model fits faster. With Stage A and P7 this places the class signal in a learned low-dimensional linear subspace, as the score-contrast space of a discriminant does, not in the leading principal components. Neither result shows that more than twelve learned features are needed, nor that a spatial model leaves that information unused (with natural 3×3 context the encoder supports comparator-level recovery, P9).

Table 19: Experiment 4, non-denoised copy, fold 0, validation macro-F1. (a) Criteria of Table 18(c), for P49 and P50(a) in every seed and in the mean; the budget-end verdicts are there. (b) Donor pools (CancerEpi / CAS / NormalStroma / NormalEpi): validation 1400 / 1200 / 600 / 500, test 1000 / 1000 / 500 / 200; residual same-core donors 0 (validation), 0.0013 / 0.0013 / 0.0008 (test, seeds 0 / 1 / 2). $\Delta_{\mathrm{cm}} = \mathrm{iid} - $ class-matched, $\Delta_{\mathrm{hom}} = \mathrm{iid} - $ homogeneous (paired); $\Delta_{\mathrm{cm}}(L^*)$: secondary reading; R: the comparator's best registered condition. Registered, every seed: P58, $\Delta_{\mathrm{cm}} \leq 0.03$ (met, $3/3$); P59, $\Delta_{\mathrm{cm}} > 0.10$ (not met, $0/3$); P60, the validity control, $|\mathrm{cm}(\mathrm{shuf}) - R| \leq 0.05$ (met, $3/3$); $\Delta_{\mathrm{cm}} - \Delta_{\mathrm{hom}}$ has no threshold. Archived: per-arm values at the first crossing of L^*, 400,000-step budget-end rows, test columns, comparator scores, retrained heads.

(a) Matched fit $L^* = 0.30$ under a tenfold step ceiling (registration 27.4): clause readings

P49 (P7)	not met (0.064)	P50(c) (P11, clause 2)	met (0.025; new fit point)
P50(a) (P8, clause 2)	met (0.682; seed 0 by 0.0007)	P51 (P10)	not evaluable ($0/3$)
P50(b) (P9)	head-level; encoder-level not met	secondary (rate)	not met
all-arm rule	not evaluable (frozen arm not extended)		

(b) Class-matched incoherent context (registration 27.7), natural-context arm, budget end, per seed and mean

seed	iid	ctx-rnd	homog.	class-matched	Δ_{cm}	Δ_{hom}	$\Delta_{\mathrm{cm}} - \Delta_{\mathrm{hom}}$	$\Delta_{\mathrm{cm}}(L^*)$	$\lvert\mathrm{cm}(\mathrm{shuf}) - R\rvert$
0	0.733	0.234	0.704	0.843	−0.110	+0.029	−0.139	−0.139	0.024
1	0.741	0.237	0.708	0.865	−0.124	+0.032	−0.156	−0.164	0.026
2	0.723	0.232	0.702	0.833	−0.110	+0.021	−0.131	−0.128	0.029
mean	0.732	0.234	0.705	0.847	−0.115	+0.027	−0.142	−0.144	–

E.3 Fold-0 diagnostics and five partitions (non-denoised copy)

Matched fit under a tenfold ceiling (registration 27.4). The four arms with a trained encoder that ended the budget without an $L^* = 0.30$ checkpoint (natural 16-PC, shuffled, shuffled 16-PC, $\kappa = 1/256$) were run again with the step ceiling raised tenfold to $400{,}000$ and nothing else changed; an identity gate reproduced all twelve runs exactly up to step $40{,}000$ (readings in Table 19(a)). The frozen arm was not extended, so the registered all-arm matched-fit rule remains not evaluable and the endpoint comparison stays primary. Three arms reach L^* in every seed, after $57{,}302$–$235{,}430$ steps against $12{,}684$–$13{,}943$ for the natural-context arm: matched fit, not matched time. The shuffled 16-PC arm stays censored at the tenfold ceiling (training loss 0.317–0.326), a persistent non-attainment under the tested optimizer and budget, not an upper bound on capacity or performance; the redundancy reading of P9 is not supported under any outcome because G1 passed. On this copy the comparisons attaining loss 0.30 retain the head-level recovery reading and a 0.064 macro-F1 gap after PCA-16; the shuffled PCA-16 arm remains censored after $400{,}000$ steps, so neither an all-arm matched-fit conclusion nor a capacity limit is established.

A label-conditioned diagnostic: class-matched context (registration 27.7; Table 19(b)). Context-random evaluation changes two things at once: the neighbours stop sharing the centre's class, and the patch stops being a coherent neighbourhood. A fourth condition tests whether the drop persists when class information is preserved: each of the eight neighbours is replaced by the centre spectrum of a pixel from another core selected using the centre's ground-truth class, an evaluation diagnostic and not a deployable intervention (donors drawn uniformly within the class, redrawn if they share the centre's core; pools and residual same-core fractions in Table 19). The natural-context and shuffled arms were trained again at the unchanged budget with this condition added; an identity gate reproduced all six runs exactly. Replacing the neighbours by spectra from other cores selected using the centre's ground-truth class yields mean validation macro-F1 0.847 ($\Delta = \text{iid} - \text{class-matched} = -0.110$, -0.124 and -0.110 in seeds 0–2, budget end), versus 0.732 with natural context and 0.234 with context-random donors. Thus destroying the original spatial arrangement while preserving neighbour class information does not reproduce the context-random collapse in this fold and model. The shuffled comparator, trained on uninformative context, scores within 0.05 of its best registered condition in every seed and so passes the registered validity check; this does not establish absence of all confounding. At L^* the same reading holds in every seed. The homogeneous condition, which also keeps the class and destroys the arrangement, costs the natural-context model 0.021–0.032 and does not distinguish reading the centre from reading class-consistent neighbours; the difference between the two conditions carries no registered threshold. The result concerns this dataset, fold and model: it does not show that the head uses only the neighbours' class, and it is not carried over to Pavia or to the other folds. The registration was informed by a 100-step, one-seed pilot of a different arm that pointed the same way (disclosed in registration 27.7.11); its bands were not taken from the pilot and were not changed after it.

Table 20: Experiment 4, **non-denoised copy only**: registered verdicts on the five patient partitions (fold 0^* registered; seeds 0–2; criteria of Table 18(c); G1 on validation *and* test for all three families, Stage B on validation at budget end, P9 with the budget-end encoder; $O(\cdot)$: MLP in folds 0–2, LR in 3–4; secondary: sign of the fold-mean difference, $\kappa = 1/256$ minus nat). In every fold `shuf16` and `frozen` reach no registered threshold, so the registered matched-fit comparison is unavailable. Numeric rows archived.

	G1	P8, cl. 1	P8, cl. 2	P9, head	P9, encoder	P11, cl. 1	P11, cl. 2	P7	sec., probe	sec., recovery
folds met	5/5	5/5	3/5	5/5	0/5	4/5	4/5	0/5	0/5	1/5
not met in	–	–	folds 2, 3	–	all	fold 4	fold 4	all	all	all but fold 1

Five patient partitions (Table 20). The gate, clause 1 of P8 and the head-level reading of P9 hold in $5/5$ folds and P7 fails in all five; clause 2 of P8 holds in $3/5$ (short by 0.012 and 0.030 in folds 2 and 3) and P11 in $4/5$ (fold 4 misses both tolerances, at 0.026 and 0.031), so two registered clauses are fold-dependent. In every fold the natural 16-PC, shuffled and $\kappa = 1/256$ arms end the budget without an $L^* = 0.30$ checkpoint, so their L^* retraining cells are censored (45 of 135); the tenfold ceiling and the class-matched condition above cover fold 0 only. The folds share cohort and training cores, and the extension covers Experiment 3's constructed setting and this copy only, not the denoised copy, the public scene or any other measurement.

E.4 A REGISTERED SCREEN FOR NATURAL 15×15 NEIGHBOURHOODS

A prespecified 15×15 natural-annulus screen failed its readiness gate under the primary whitening (annulus readout 0.41–0.42 against the required 0.60; $0/3$; Table 21), so no joint-training experiment was launched. The registered standardized secondary analysis passed ($3/3$); this preprocessing dependence is reported without changing the primary decision.

Table 21: The 15×15 annulus screen (registration amendment 21.1b): macro-F1 of ring-mean logistic readouts on held-out fold-0 training cores, random-encoder seeds 0–2 (majority baseline 0.135; permutation $p_{95} \leq 0.263$). The gate requires the random-encoder annulus readout to reach 0.60 and the trained centre reference 0.65, each exceeding the random-encoder centre readout by 0.10.

preprocessing	random-enc. centre	random-enc. annulus	trained centre (942-dim)	gate
whitened (primary)	0.31 / 0.33 / 0.34	0.41 / 0.42 / 0.41	0.75	fail, $0/3$
standardized (secondary)	0.54 / 0.60 / 0.60	0.73 / 0.78 / 0.73	0.79	pass, $3/3$

F PUBLIC HYPERSPECTRAL REPLICATIONS (PAVIA UNIVERSITY AND SALINAS)

Our tissue spectra are not public, so Experiments 3 and 4 were repeated on the ROSIS Pavia University scene (610×340 pixels, 103 bands, nine classes, $42{,}776$ labelled pixels) with the same code, protocols, arms and the predictions of registration section 9, recorded by the authors before the first run and first committed with the code and first results (Appendix H; the readiness scan and the per-pixel sweep are descriptive stages, as on tissue). The scene is cut into 20×20-pixel tiles assigned at random (seed 0) as $0.6/0.2/0.2$ ($208/70/70$ labelled tiles, ≥ 60 admissible pixels per class and split); a pixel is admissible only if its 3×3 neighbourhood lies inside its tile, so no evaluation neighbourhood overlaps a training one, and up to $400/200/200$ pixels per tile and class are sampled ($19{,}870/6{,}172/6{,}407$ of $34{,}386$ admissible). Residual autocorrelation across tile borders is not removed: this is a tile split within one scene, not an independent-scene replication. Standardization and PCA are fitted on the training split, with no per-pixel normalization, denoising or derivative.

Table 22: Pavia, constructed context (Meadows vs Trees): binary accuracy at $L^* = 0.30$, validation unless marked (majority baseline 0.784, test 0.838), mean±sd, $n = 3$; no cell censored; indented rows per seed. Headers: readiness scan on tile-disjoint training-split halves (centre-only oracle; random $K = 12$ encoder, mean [min, max] of three; $|\cos| = 0.91$ with the first principal direction), initial probe per seed. Retr. rev.: fresh head (Appendix C.2) on the arm's frozen encoder (`frozen`: initial). Other pairs, widths and $\kappa = 1/16$ in full, per-seed recovery: archived.

arm	M / κ	step	probe gain	rev.	ctx-rand.	rev. (test)	retr. rev.
High accessibility (standardized): oracle 0.933; random encoder 0.850 [0.84, 0.88]; initial probe $0.915, 0.923, 0.909$							
`sp`	32	1066±262	-0.001±0.004	0.230±0.021	0.544±0.010	0.227±0.027	0.887±0.010
`headlr`	32 / $\frac{1}{256}$	13524±4802	0.003±0.004	0.189±0.028	0.512±0.041	0.181±0.043	0.907±0.004
slow−fast by seed: probe $-0.001, +0.006, +0.008$; retr. rev. $+0.014, +0.020, +0.027$							
`ctxfree`	128	4781±463	0.003±0.002	0.875±0.002	0.877±0.006	0.964±0.002	–
`frozen`	32	1418±232	0.000±0.000	0.217±0.016	0.542±0.005	0.213±0.016	0.885±0.010
Low accessibility (whitened), $\gamma = 30$: oracle 0.955; random encoder 0.667 [0.62, 0.70]; initial probe $0.664, 0.703, 0.755$							
`sp`	32	42±5	-0.001±0.002	0.071±0.022	0.476±0.024	0.066±0.024	0.658±0.018
`sp` at every width ($M = 8, 32, 128, 512$), per seed: $12/12$ reach L^*; probe gain -0.004 to $+0.004$; 13 to 158 steps							
`headlr`	32 / $\frac{1}{256}$	250±23	0.002±0.020	0.080±0.016	0.472±0.029	0.072±0.025	0.649±0.003
slow−fast by seed: probe $+0.023, -0.009, -0.006$; retr. rev. $+0.012, -0.021, -0.019$							
`ctxfree`	128	9142±764	0.223±0.033	0.851±0.015	0.857±0.017	0.872±0.010	–
`frozen`	32	53±8	0.000±0.000	0.072±0.023	0.476±0.023	0.067±0.024	0.660±0.024
Low accessibility (whitened), $\gamma = 10$ *(amendment)*: scan and initial probes as for $\gamma = 30$							
`sp`	32	283±28	0.004±0.005	0.054±0.012	0.486±0.011	0.052±0.016	0.687±0.024
`sp` at every width ($M = 8, 32, 128, 512$), per seed: $12/12$ reach L^*; probe gain -0.002 to $+0.018$; 88 to 662 steps							
`headlr`	32 / $\frac{1}{256}$	1901±96	0.029±0.046	0.059±0.006	0.479±0.015	0.056±0.014	0.697±0.026
slow−fast by seed: probe $+0.073, -0.002, +0.005$; retr. rev. $+0.061, -0.032, +0.001$							
`ctxfree`	128	8923±880	0.231±0.037	0.899±0.005	0.900±0.011	0.920±0.003	–
`frozen`	32	368±42	0.000±0.000	0.054±0.011	0.484±0.009	0.053±0.014	0.687±0.036

Experiment 3: constructed context. The pair is Meadows ($+1$) against Trees (-1), the two largest vegetation classes and a standard confusion in this scene ($8{,}347$ and $1{,}322$ admissible training pixels; readiness scans for three additional pairs are archived). The readiness scan reproduces the tissue structure in both preprocessings with a smaller asymmetry, so both are kept as accessibility regimes ($\tau = 1.89$ standardized, 1.66 whitened). Up to $2{,}400$ balanced training patches and 400 probe-fit pixels are used and all validation and test pixels of the pair are evaluated; everything else, recovery included (Appendix C.2), is as in Appendix D. All 90 regime/arm/width/rate/seed cells reach $L^* = 0.30$ (81 distinct runs: the $M = 32$ baseline and the $\kappa = 1$ arm are one run), so none is censored. Training patches are class-balanced, but accuracies are read against the majority baseline and paired within-seed differences carry the comparisons.

In both whitened (low-accessibility) regimes suppression and reliance replicate: the standard-parameterization arms leave the probe within 0.019 of its initial value across the tested widths and fit in tens to hundreds of steps (Table 22, every-width rows), while the uninformative-context comparator raises the probe by 0.22 to 0.24 in about $9{,}000$ steps, and the informative-context arms fail under reversal. In the standardized (high-accessibility) regime reliance replicates without a detected probe deficit (comparator probe gain 0.003): the informative-context arms fail under reversal (0.19 to 0.23; comparator 0.875), and a fresh head recovers 0.875 to 0.911 from every encoder. *Both registered*

Table 23: Pavia, natural context: nine-class macro-F1 on validation, mean±sd, $n = 3$ seeds. Stage B at budget end (40,000 steps, every run); *reached*: seeds reaching training loss $0.5/0.4/0.3$. Gate G1: $k = 16$ components against all 103 bands. Recovery: fresh head on the frozen encoder (Appendix C.2); the random-encoder row is the empirical reference (macro-F1 has no single chance level). Archived: test split, matched-loss rows, k-sweep, homogeneous column.

arm	loss	reached	probe	probe gain	iid	ctx-random	n
natural, no bottleneck	0.21±0.01	3/3/3	0.790±0.002	0.000±0.000	0.836±0.006	0.152±0.008	3
natural	0.21±0.01	3/3/3	0.732±0.015	0.017±0.011	0.840±0.004	0.144±0.016	3
natural, $\kappa = 1/256$	0.84±0.06	0/0/0	0.702±0.015	-0.013±0.010	0.512±0.016	0.090±0.007	3
shuffled, no bottleneck	0.47±0.01	3/0/0	0.790±0.002	0.000±0.000	0.648±0.011	0.654±0.004	3
shuffled (comparator)	0.49±0.01	3/0/0	0.722±0.011	0.006±0.004	0.594±0.049	0.627±0.008	3
natural, frozen encoder	0.41±0.01	3/0/0	0.715±0.007	0.000±0.000	0.703±0.013	0.169±0.006	3

Stage A gate G1

classifier	$k = 16$	$k = 103$	Δ	Δ (test)
LDA	0.767	0.781	+0.014	+0.017
LR	0.816	0.830	+0.014	+0.011
MLP	0.891±0.004	0.911±0.005	+0.020	+0.009

Fresh-head recovery

encoder, saved	ctx-random	iid
natural, random init	0.559±0.004	0.547±0.025
natural, at $L = 0.3$	0.619±0.004	0.624±0.008
natural, at end	0.629±0.005	0.637±0.014
$\kappa = 1/256$, at end	0.620±0.010	0.619±0.041
shuffled, at end	0.615±0.005	0.602±0.049

Table 24: Registered batch K6, tissue (non-denoised copy) against Pavia: per-seed validation macro-F1 differences from per-pixel retraining on the frozen natural-context encoder at budget end. One-token head: centre-pixel output, $M = 32$ ReLU head, plain gradient descent, 20,000 steps, every cell budget-limited (loss still falling by more than 1% over the last ≈ 2,000 steps); LR: converged logistic regression. P55, P55L: met if every seed ≥ -0.05, shortfall if every seed < -0.10. Each macro-F1 entering a difference is a lower bound on decodable accuracy under its own model family, protocol and budget.

clause	difference; criterion	tissue	verdict	Pavia	verdict
P55	one-token head − raw-band MLP	$-0.011, -0.010, -0.016$	met	$-0.272, -0.233, -0.247$	budget-limited shortfall
P55L	LR on the encoder − LR on raw bands	$+0.010, +0.015, +0.004$	met	$-0.106, -0.098, -0.094$	not established
P56	one-token − retrained nine-token head; > 0	$+0.030, +0.002, +0.029$	met	$+0.018, +0.018, +0.031$	met
P57	LR, natural − shuffled encoder; ≥ -0.02	$+0.005, +0.004, -0.002$	met	$+0.018, +0.036, +0.024$	met

rate predictions fail: they required the within-seed probe change from $\kappa = 1$ to $1/256$ to be positive in every seed, and at neither amplitude is it (slow−fast rows; the largest response is in the seed with the least accessible initialization); reversal accuracy is unchanged. The paired slow−fast recovery difference changes sign between seeds in both low-accessibility regimes; in the high-accessibility regime the slow encoder recovers best in every seed by a small margin, an ordering not shown to be mediated by the probe, whose within-seed changes have a different sign pattern. The whitened Pavia construction starts more accessible than the breast one and its informative-context fits are rapid, which may affect sensitivity to the module-rate intervention but does not explain the inconsistency: context-random training still improves the probe substantially and slowing the head does not reliably recover that improvement, so the failed rate prediction is a limit of generalization, not an identified consequence of insufficient room for adaptation.

Experiment 4: natural context. Nine classes and macro-F1, not comparable with the binary accuracies above, with the Stage A gate and the Stage B arms of Appendix E (Table 23). Stage A gives the opposite of the tissue result: the best per-pixel classifier gains only 0.020 macro-F1 on validation and 0.009 on test beyond the sixteenth component, so the registered gate fails — a performance result for these classifiers, not an information bound (the top sixteen components carry 99.93% of the standardized training variance, against 81.7% on tissue: a description of the inputs, not a demonstrated cause) — and by the registered rule the compression arms, and with them P7 and P10, were not run. In Stage B 88% of neighbours share the centre label, 12% are unlabelled and none disagrees, as on tissue; P8's first clause is met, P9 reads head-level and P11 is met. No accessibility deficit is detected (probe 0.732 against the comparator's 0.722, from 0.715 at initialization; standardized, this scene is in the high-accessibility regime), which is not evidence that the encoder retained everything. Two limitations of the head family are visible: the $M = 32$ ReLU head trained

by plain gradient descent is a weak nine-class per-pixel classifier (comparator 0.627, no-bottleneck comparator 0.654, per-pixel MLP 0.911), so P8's second clause fails, without separating whether the architecture or the optimizer is responsible; and the slow-head arm ($\kappa = 1/256$) reaches no matched-fit threshold within the budget, so the rate prediction is not evaluable at matched fit and is not met at budget end.

Per-pixel retraining on the frozen encoder (registered batch K6, Table 24). On tissue, centre-only retraining from the fixed natural-context encoder reaches within 0.02 of the raw-spectrum MLP in every seed and the linear comparison meets its registered tolerance. On Pavia the one-token head falls 0.23 to 0.27 short of the per-pixel MLP with all three cells budget-limited; the registration reports this as a shortfall at $20{,}000$ steps and licenses no attribution from it, and the linear comparison is not established either way. On both datasets the one-token head beats the nine-token patch head retrained on the same encoder and budget in every seed (P56), which shows an improvable head or protocol component at a fixed encoder, not that the whole patch-head deficit is unrelated to the encoder (on Pavia the remaining gap to the raw-spectrum references is unresolved); and training with informative context met the registered retention tolerance relative to the comparator at the end of training (P57).

What replicated and what did not. Constructed low-accessibility settings replicate a probe deficit relative to decorrelated-context training and reliance under reversal. The high-accessibility setting replicates reliance without a detected probe deficit, with strong recovery from every encoder. The low-accessibility rate and recovery responses are inconsistent across seeds, and the registered rate prediction is not met at either amplitude. Natural-context Pavia has the procedure-specific head-level reading, but its weak comparator and failed compression gate limit what it tests, and the whole replication is a tile split within a single scene.

F.1 Salinas: a second public scene (registered batch K10)

Experiments 3 and 4 were repeated on a second scene, AVIRIS Salinas (512×217 pixels, 204 bands, sixteen classes, $54{,}129$ labelled pixels) with Pavia's code, split rule, preprocessing, arms, seeds, L^*, censoring and disposition, registered with predictions P67–P77 on 21 September 2026 (registration section 27.10) before any registered run; the 20×20 tile split keeps all sixteen classes (feasible at attempt 277; $114/38/38$ tiles), under Pavia's single-scene caveat. Among the four pairs of a descriptive readiness scan on tile-disjoint halves of the training split, the two-regime rule used for tissue and Pavia selects only Soil_vinyard_develop against Corn_senesced_green_weeds (centre-only oracle 0.984 standardized, 0.986 whitened; a random 12-dimensional encoder reads it at 0.973 and 0.550, the lowest whitened value scanned), fixed after the scan and before any registered training run. Its contrast is strongly aligned with the first principal direction ($|\cos| = 0.91$), so Salinas does not test the off-principal-direction setting; the one off-direction pair scanned (Grapes_untrained against Vinyard_untrained, $|\cos| = 0.41$, oracles 0.718 and 0.728) was screened out by the rule, not shown incapable of supporting any useful experiment. A check of the registered pre-run facts (including $\tau = 1.298$, 1.307, 1.314) passes $62/62$.

Constructed context (Table 25). Suppression and reliance at $\gamma = 30$ (P67, P68) and both controls (P70, P72) are met; P72 makes the ready-regime reversal failure head-level, not encoder suppression. The rate prediction P69 is not met (its reversal clause is). P71 names no regime, so all three are reported: its first clause holds at $\gamma = 10$ only, whereas Pavia met it in the ready regime only; its second clause holds in every cell.

Natural context (Table 26). P73–P77 re-register P7–P11 (Appendix E). As on Pavia, the failed MLP gate leaves the sixteen-component reading uncontradicted — a performance result, not an information bound — and P73 and P76 were not run; the three arms that never reach L^* are compared only at the $40{,}000$-step endpoint. P74 meets its first clause and not its second. P75 reads head-level: the fresh head on the natural-context encoder frozen at budget end reaches the registered bar at that labelled endpoint, not at matched loss, and as recovery of the registered comparator level only, not of everything the full-band MLP uses. P77 meets its first clause and not its second, so the learned bottleneck is not shown to lose nothing; on Pavia both clauses were met. The secondary slow-head clause is not met.

Table 25: Salinas, constructed context: registered clause-level verdicts (binary accuracy on validation at $L^* = 0.30$, seeds 0–2; Soil_vinyard_develop $+1$ vs Corn_senesced_green_weeds -1; majority baseline 0.597 validation, 0.818 test; all 90 runs reach L^*). Recovery (Appendix C.2) retrains a fresh head on the encoders saved at loss 0.30; unlike there, the ready-regime retraining runs stop earlier, at training loss 0.10; slow−fast: paired retrained reversal accuracy, $\kappa = 1/256$ minus $\kappa = 1$. Per-seed values and long tables are archived.

Clause	Registered criterion	Observed	Verdict
P67 suppression	$\gamma = 30$: `sp` probe gain $<$ `ctxfree`'s, $M = 8, 128$, every seed	$6/6$; margins 0.272–0.361	met
P68 reliance	$\gamma = 30$ rev.: `sp` ≤ 0.5 (every M, seed); `ctxfree` ≥ 0.75	$12/12$ (max 0.111); $6/6$ (min 0.821)	met
P69 rate, 1	$\gamma = 30$, $M = 32$: probe gain rises strictly, $\kappa = 1 \to 1/16 \to 1/256$, every seed	$1/3$ seeds	not met
P69 rate, 2	mean reversal spread across $\kappa < 0.05$	$0.063, 0.058, 0.059$ (0.005)	met
P70 reference	`frozen` reversal < 0.5, every regime and seed	$9/9$ (max 0.228)	met
P71 recovery, 1	slow−fast > 0, every seed	$\gamma = 10$: $3/3$; $\gamma = 30$: $1/3$; ready: $1/3$	$\gamma = 10$ only
P71 recovery, 2	retrained rev. from the $\kappa = 1$ encoder $\leq$ from the random one $+ 0.05$	$9/9$ (largest excess -0.021)	met
P72 control (ready)	$\lvert$probe gain$\rvert \leq 0.05$ at every κ	$9/9$ (max 0.024)	met

Table 26: Salinas natural-context experiment; sixteen-class macro-F1 on validation, registered clause-level verdicts; mean±sd over three seeds. Stage A compares 16 principal components with all 204 bands. Stage B values are at the 40,000-step endpoint. Recovery (Appendix C.2) freezes the encoder at the labelled checkpoint; the random-encoder recovery is an empirical reference, not a theoretical chance level; $O = 0.954$ is the full-band MLP. Long tables are archived.

Clause	Registered criterion	Observed	Verdict
Stage A gate	MLP gain ≥ 0.02, $k = 16 \to 204$ bands, val. and test	MLP $+0.003$, $+0.012$; LDA $+0.022$, $+0.024$, LR -0.005, $+0.010$ (descriptive)	failed
P73, P76	compression arms; require the gate	not run	not run
L^* attainment	censored arms compared at budget end only, never at matched fit	shuffled, frozen, slow head $0/3$; others $3/3$	3 censored
P74, clause 1	ctx-random, natural $\leq O - 0.10 = 0.854$	0.073±0.002 (at L^*: 0.075±0.006)	met
P74, clause 2	ctx-random, shuffled $\geq O - 0.03 = 0.924$	0.786±0.005	not met
P75, head-level	recovery, natural (end) $\geq$ shuffled $- 0.03 = 0.756$	0.843±0.011 (random init 0.577±0.046; shuffled encoder 0.780±0.010)	obtains
P75, encoder-level	recovery $<$ shuffled -0.03 and shuffled $\geq O - 0.03$ (≥ 0.924)	neither	does not
P77, clause 1	$\lvert$iid: natural − no bottleneck$\rvert < 0.02$	0.944±0.007 vs 0.949±0.004: 0.004	met
P77, clause 2	$\lvert$ctx-random: shuffled − no bottleneck$\rvert < 0.03$	0.786±0.005 vs 0.895±0.011: 0.109	not met
Secondary	$\kappa = 1/256$ above $\kappa = 1$, probe gain and recovery, every seed	probe $2/3$, recovery $0/3$	not met

What Salinas adds. Clause by clause the outcomes are mixed: at $\gamma = 30$ a probe deficit relative to decorrelated-context training and reliance under reversal replicate on a second scene, but a useful rate intervention is a different claim — the registered $\gamma = 30$ rate prediction fails on Salinas as on Pavia, and recovery ordering depends on the scene and the amplitude. Because Salinas starts with a lower whitened random-encoder probe than Pavia (0.550 against 0.667 at $K = 12$) and the rate prediction still fails, the earlier explanation that Pavia had too little accessibility headroom is insufficient on its own; we infer no new cause from this comparison.

F.2 HeiPorSPECTRAL: a third public dataset, porcine visible/near-infrared (registered extension)

In an animal-disjoint HeiPorSPECTRAL extension (Studier-Fischer et al., 2023), we used constructed $\gamma = 30$ contexts around measured liver and gallbladder pixel spectra ($M = 32$, $K = 12$, L1 normalization followed by training-fitted whitening), the pair selected by a training-only eligibility screen from three nominated organ pairs; the selected contrast is strongly aligned with the leading principal component. All fifteen arm/seed combinations reached loss 0.30 and nothing was censored. Suppression and reliance replicated, while the registered rate-response and recovery-ordering predictions were not met. Blocking at the slow head rate gave mean matched-loss probe and recovery differences of $+0.015$ and $+0.018$, below the registered 0.10 requirement and with one negative

seed in each comparison; the separate interaction sign criteria were met. The registered equal-update criteria were not met at either $4{,}000$ or $40{,}000$ updates, with mean recovery differences -0.048 and -0.026. The blocked slow arm reached L^* after $5{,}415$–$12{,}658$ updates against 124–164 for the open slow arm (44–$77\times$), so the matched-loss comparison does not equalize update counts. These results concern one PC1-aligned pair, three seeds, five contributing training animals and two validation plus two test animals; the contexts are constructed, not natural neighbourhoods, and the data are porcine organs, not human histology. Donors are sampled without consulting the centre label, from the same split, with one redraw when the donor and centre animals coincide. Table 27 gives every clause; the registration, report, per-animal tables, provenance (member-level checksums of the selectively fetched official archive) and gate logs are in the release.

Table 27: HeiPorSPECTRAL extension (registered on its branch before the run; predictions HP1–HP9, 23 clauses): validation animals P095 and P096 pooled, seeds 0–2 paired within seed; one row per clause and metric. Arms: A1 open $\kappa = 1$, A2 open $\kappa = 1/256$, A3 blocked $\kappa = 1$, A4 blocked $\kappa = 1/256$, A5 context-free comparator. Probe gain: LDA probe minus the shared random-encoder probe at matched loss $L^* = 0.30$; rev: the joint classifier's reversed-context accuracy; recovery: fresh-head reversed accuracy after $20{,}000$ steps (Appendix C.2); D, R: blocked-minus-open probe and recovery at equal updates (A4 − A2); interaction $I = (\text{A4} - \text{A2}) - (\text{A3} - \text{A1})$ for each metric. HP7 requires $D > 0$ in every seed and mean $D \geq 0.10$; HP8 requires $R \geq 0.10$ in every seed and in the mean (both sub-clauses failed wherever printed as one row). HP4 requires a mean of at least 0.10 and a positive difference in every seed, HP5 only a positive mean with two of three positive seeds, which is why HP4 and HP5 differ. Values to 3 dp, 4 dp where 3 dp would print a signed zero. No cell censored; all fifteen arm/seed runs reached L^*; all 42 recovery runs completed. Test animals are descriptive and enter no verdict.

clause, quantity	criterion	s0	s1	s2	mean	verdict
HP1 suppression, probe gain A1 − A5	< 0 every seed	-0.300	-0.278	-0.335	-0.304	*met*
HP2a reliance, rev(A1)	≤ 0.5 every seed	+0.094	+0.080	+0.102	+0.092	*met*
HP2b comparator, rev(A5)	≥ 0.75 every seed	+0.909	+0.890	+0.863	+0.887	*met*
HP3a rate response, probe gain A2 − A1	> 0 every seed	-0.004	+0.007	+0.004	+0.003	*not met*
HP3b reversal change A2 − A1	$\lvert\text{mean}\rvert < 0.05$	+0.006	+0.008	-0.003	+0.003	*met*
HP4a blocking at L^*, probe gain A4 − A2	mean ≥ 0.10, > 0 every seed	+0.030	-0.0005	+0.015	+0.015	*not met*
HP4b blocking at L^*, recovery A4 − A2	mean ≥ 0.10, > 0 every seed	+0.039	-0.011	+0.025	+0.018	*not met*
HP5a interaction I at L^*, probe	mean > 0, > 0 in $\geq 2/3$	+0.027	+0.001	+0.018	+0.015	*met*
HP5b interaction I at L^*, recovery	mean > 0, > 0 in $\geq 2/3$	+0.035	-0.007	+0.027	+0.018	*met*
HP6a recovery ordering A2 − A1	> 0 every seed	-0.007	+0.012	+0.003	+0.002	*not met*
HP6b recovery A1 − init	mean ≤ 0.05	-0.002	+0.003	+0.002	+0.001	*met*
HP7 D at 40,000 updates, probe	> 0 every seed; mean ≥ 0.10	+0.052	-0.070	-0.009	-0.009	*not met*
HP8 R at 40,000 updates, recovery	≥ 0.10 every seed and mean	+0.021	-0.077	-0.024	-0.026	*not met*
HP9a I at 40,000, probe	mean > 0, > 0 in $\geq 2/3$	+0.059	-0.025	-0.002	+0.010	*not met*
HP9b I at 40,000, recovery	mean > 0, > 0 in $\geq 2/3$	+0.027	-0.029	-0.013	-0.005	*not met*
HP7b D at 4,000, probe	as HP7	+0.009	-0.051	-0.061	-0.034	*not met*
HP8b R at 4,000, recovery	as HP8	-0.020	-0.065	-0.059	-0.048	*not met*
HP9b(a) I at 4,000, probe	as HP9	+0.018	-0.014	-0.049	-0.015	*not met*
HP9b(b) I at 4,000, recovery	as HP9	-0.019	-0.018	-0.045	-0.028	*not met*

G SCOPE LIMITATIONS FROM A PRODUCTION MODEL

An earlier diagnostic phase studied the production model of O'Leary et al. (2026) (a linear per-pixel map $\mathbb{R}^{942} \to \mathbb{R}^K$ into a vision transformer of width M and a convolutional segmentation head) on breast QCL cores; $M \in \{48, 192, 384\}$, three seeds, fold 0. Its retraining runs do not demonstrate the mechanism of Theorem 1, which they do not measure (its head hypothesis is unmet; no arm isolates a two-timescale flow): an apparent freezing advantage there is evidence neither for nor against it. What Section 1 cites is that its curvature width trend is normalization-sensitive. At initialization the curvature disparity $\mathcal{D}_{\text{curv}}$ (spatial over spectral top Gauss–Newton eigenvalue) is 0.33 ± 0.12, 0.46 ± 0.21, 0.70 ± 0.19 at $M = 48, 192, 384$, at or below parity at every buildable width (removing the spectral normalization deflates the denominator $2.4\times$ and crosses parity): only a trend is at issue, not a large disparity. Freezing both head batch normalizations at their initial running statistics removes it (pooled log–log width slope $+0.613 \to +0.006$ for the first head convolution,

$+0.398 \to -0.331$ for the spatial block), the incoming features unchanged. For $c > 0$, scaling that convolution's weight and bias by c and the following train-mode normalization's ϵ by c^2 preserves the logits and scales its top curvature by c^{-2} (with ϵ fixed, as implemented, only approximately): the curvature is a parameterization quantity, not an architectural one (van Laarhoven, 2017; Karakida et al., 2019). The only conclusion drawn is that it is not a usable diagnostic here.

Registered outcomes (historical). Fine-tuning a pretrained encoder falling below freezing it: *not supported*; in twenty-one runs at $M = 48$ that equalize the original protocol's clipping, affine-parameter and checkpoint asymmetries, neither matched pair resolves at $n = 3$ (pre-registered final-window 90% paired-t intervals: joint-linear minus frozen-random $[-0.142, +0.163]$, fine-tuned minus frozen-pretrained $[-0.022, +0.063]$). Harmful-drift ordering $\times 0.1 \geq \times 1 \geq \times 10$ in the spectral learning rate: *not observed*; $\times 10$ beats $\times 1$ on (exploratory) best-validation in $3/3$ seeds, while its final-window endpoint is the lowest and most variable (0.561 ± 0.114). Scope: one validation fold, $n = 3$, no test set; the original 39-run protocol confounds its joint-versus-frozen contrast, saved no checkpoint (its best-validation is a curve maximum at which no saved checkpoint is available) and its two selection policies order the arms oppositely; the pretrained encoders were selected on the *same* cores later used to score them; the pairs are not established as bit-identically matched. Contrast table, schedule and normalization-statistics diagnostics and full account: `docs/appendix_archive/G_*.tex`.

H VERIFICATION RECORD

Each row of Table 3 (the checks of Appendix A) and of Table 28 below is a script or built-in self-test in the accompanying repository, with fixed seeds and a recorded outcome, and certifies nothing beyond its statement; the repository's README maps each row to its script and exact command and archives the printed output and per-check detail.

Table 28: Implementation checks; *outcome* includes what is deliberately not tested.

Check (claim)	Coverage	Outcome
Head implementation against the library convolution, self-test (Exp. 2)	$M = 2, 37, 512$, fp64	Agree to 10^{-15}
Route-block identity, self-test (C8)	Gate, fp32 and fp64	Max $\lvert\Delta\rvert = 0$
Accessibility preparation, self-test (C2)	Gate, all 24 runs	Worst 3.9×10^{-9} against 10^{-3}
Extension identity gates in the table generators (C8b, C2b, C11b, C12)	Shared snapshots, 1,600 rows	PASS in all four, max difference 0; failure voids the extension
Held-out discipline (Exp. 3); registrations (all predictions)	Readiness scan; the registration record and its commit history	Training patients only, core-disjoint; registrations recorded before their runs (commit-verified from §21 on; earlier sections by the authors' record, see below), plus explicitly labelled retrospective (C8c), post-outcome (§4.4c; Exp. 4 budget end, §8.6) and outcome-prompted ($\gamma = 10$, §5; K1, C2b, C11b) analyses; censoring recorded

Registration times. From section 21 (16 September) on, a registration's time is the commit time of its section and no batch starts before that commit (checked against the launch logs of C8b and C2b). The clock times written into sections 4–12 and 17–19 (10–14 September) are hand-written and run ahead of the machine clock by a drifting offset of half an hour to fifteen hours, so they cannot be compared with commit or file times; the pre-registration file carries a dated correction (§27.13) listing, per section, the first commit containing its text and the first result time. Where a commit or launch log precedes the first output, the machine record supports the recorded order: the $\gamma = 10$ amendment, section 17, the slow-arm and recovery outcomes of section 19, the P15 retraining of section 10, and seven of the eight fold blocks of section 12. Where the text was first committed with or shortly after its code and first results — section 6b's step budget and stopping rule, sections 8 and 9, the section-10 grid, the first seed-3/4 runs of section 11 and Experiment 3's fold 1 — the order rests on the authors' record: those commits were batched and show neither order.

The repository also keeps, unmodified, sixteen audits of an earlier, retired framework on which nothing here depends: fourteen exit zero, the audit of its share bound and width lemmas fails one of its 19 checks through its own quadrature (exact integration gives $0/300$ violations), and the audit of its curvature-floor events exits 1 on a torch-version abort in a redundant batch-normalization

re-check; both claims pass elsewhere. The diagnoses, logs and this record's earlier versions are archived (`docs/appendix_archive/H_*`).