Information-plane dynamics during learning across artificial and biological neural systems

Nikita Pospelov **Olga Ivashkina** **Viktor Plusnin** **Olga Rogozhnikova**
Anna Ivanova **Ksenia Toropova** **Konstantin Anokhin**
Institute for Advanced Brain Studies of the Moscow State University, Moscow, Russia
Correspondence: pospelov.na14@physics.msu.ru

Abstract

The information plane offers a candidate general framework for learning, defined by two axes: how much a representation preserves about its input and how much it carries about the relevant target. However, empirical evidence for information-plane dynamics in realistic-scale artificial networks and in biological systems remains limited. We address this by tracking information-plane trajectories during learning in three systems with different substrates, using a shared analytical scaffold with substrate-matched estimators. First, in a spiking ResNet-18, a biologically plausible artificial substrate gradient-trained on CIFAR-10, a two-phase fitting-then-compression trajectory emerged specifically in the deep layers; the first information-plane analysis of a spiking network at this scale. Second, on the open-source dataset of mouse visual cortex during multi-week category learning, our pipeline recovered the published cohort-level effect, now reformulated in information-plane terms. Third, in our own single-photon miniscope calcium imaging data of mouse hippocampus during multi-day place learning in open field, the axes map onto the well-established egocentric-to-allocentric transformation: the input axis becomes egocentric sensorimotor state, the relevance axis becomes allocentric position. Behavioral and decoder-baseline controls confirmed that the hippocampal trajectory reflects experience-driven learning. Across all three systems, we observed coherent motion along the relevance axis with learning, while compression, most pronounced in the deep layers of the SNN, was not universal across systems. The information plane therefore offers a substrate-independent coordinate for representation learning.

1 Introduction

Understanding how representations change with learning is central to both machine learning and systems neuroscience, yet the two fields describe this process in largely separate vocabularies: gradient flow and weight updates on one side, tuning curves and population codes on the other. The information bottleneck principle [Tishby et al., 1999, Tishby and Zaslavsky, 2015] proposes a substrate-agnostic alternative: any learned representation can be located on a two-dimensional plane defined by input retention and target relevance, and learning corresponds to motion on this plane. Building on this, Shwartz-Ziv and Tishby [2017] advanced the sharp empirical claim that representations follow a two-phase trajectory, first gaining relevance, then compressing irrelevant input information. For biological systems, this framing recasts learning as the progressive organization of experience into behaviorally relevant representations. Saxe et al. [2018] showed the original evidence depended on estimator artifacts in saturating-tanh networks. The community largely shifted toward IB-as-objective approaches [Alemi et al., 2017, Pogodin and Latham, 2020, Yang and Chen, 2023]. Another line of work uses static IB diagnostics on trained systems [Goldfeld et al., 2019, Palmer et al., 2015, Kleinman et al., 2025, Achille and Soatto, 2018]. The original question, whether compression

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

during learning is a robust phenomenon or only an artifact of particular networks, has remained untested at realistic scale and in biological systems undergoing learning.

Three recent developments make the question newly tractable. First, it has been revisited empirically: Kleinman et al. [2025] provide evidence for IB-style cortical organization during decision-making, de Mello Koch and Ghosh [2025] relate a two-phase view of IB to grokking and double descent, and Kalburge and Lengyel [2025] introduce functional IB diagnostics for probabilistic neural codes. Second, mutual-information (MI) estimators have matured [Belghazi et al., 2018, Poole et al., 2019], and compression-based methods route through autoencoder latents to avoid the binning pathology identified by Saxe et al. [2018] [Goldfeld et al., 2019, Butakov et al., 2024]. Third, chronic imaging now follows the same neurons across days in behaving animals, extending trajectory analysis from artificial networks to biological systems.

The empirical question we ask is whether the IB principle can describe learning in systems where it has not been systematically tested: biological neural circuits during learning, and a realistic-scale artificial system that provides a bridge to them. The biological systems are mouse visual cortex during multi-week supervised category learning [Goltstein et al., 2021] and mouse hippocampus during multi-day naturalistic place learning. The cortex dataset, with published behavioral and physiological learning signatures, is a methodological validation: it allows us to ask whether our pipeline recovers a known learning effect in information-plane terms. The hippocampal paradigm goes further. We apply the framework to new recordings and reformulate spatial learning in information-plane terms, with the input axis read as egocentric sensorimotor state and the relevance axis as allocentric position [O'Keefe and Dostrovsky, 1971], the two ends of the well-established egocentric-to-allocentric transformation [Wang et al., 2018, Alexander et al., 2023]. The artificial substrate is a gradient-trained spiking ResNet-18 with leaky-integrate-and-fire neurons on CIFAR-10. Discrete spike-based communication makes it a methodological bridge to biological computation [Maass, 1997]. Direct access to every neuron also enables triangulation between population-level information-plane dynamics and per-neuron tuning measurements available from Pospelov et al. [2025].

To our knowledge, information-plane trajectories during learning have not previously been tracked in a spiking network at this scale or in a biological system. Applying a shared analytical scaffold with substrate-matched estimators to all three systems, we make three empirical contributions[1]:

(i) In a gradient-trained spiking neural network (SNN) at realistic scale, a two-phase fitting-then-compression trajectory in the deep layers, and per-neuron tuning sharpness that tracked population information increasingly tightly with depth (multi-scale triangulation).

(ii) In mouse visual cortex during multi-week category learning [Goltstein et al., 2021], a published cohort-level learning effect translated from neuron-fraction terms into information-theoretic terms, together with regularities not visible at the original level of analysis: a primary visual cortex (V1) learning signal in 5/5 mice and a quantitative gap between cortical category information and behavioral readout.

(iii) In mouse hippocampus during place learning, rising relevance to allocentric position, with input compression, borderline in aggregate, concentrated in the egocentric object distances.

Across all three systems, we observed coherent motion along the relevance axis with learning, while compression, most pronounced in the deep layers of the SNN, was not universal across systems. We present the information plane as a substrate-independent coordinate for comparing representation learning across biological and artificial systems.

2 Methods: a shared information-plane scaffold

2.1 Three systems

We tracked information-plane trajectories during learning in three substrate-different systems. All datasets were analyzed with the shared scaffold of §2.2.

[1]Code (MIT license): `https://github.com/niveousdragon/ib-neurips2026-code`, release v1.0 archived at Zenodo, `https://doi.org/10.5281/zenodo.24030965`. Hippocampal recordings: DANDI:002047, `https://dandiarchive.org/dandiset/002047`. SNN activity summed over timesteps, maximally exciting inputs and learning curves (CC BY 4.0): Zenodo, `https://doi.org/10.5281/zenodo.23962331`. SNN training code [Pospelov et al., 2025]: `https://github.com/iabs-neuro/mango`.

Spiking neural network. A SpikingJelly [Fang et al., 2023] ResNet-18 with leaky-integrate-and-fire neurons (50 timesteps, binary spikes) was trained end-to-end on CIFAR-10 for 1000 epochs (learning rate 0.2) via surrogate-gradient backpropagation, reaching $\sim$86% test accuracy [Pospelov et al., 2025]. Activity was recorded as spike counts at 17 sublayers ($\mathrm{sn1}$ plus the two spike layers within each of the eight residual blocks) across 26 epoch checkpoints (epochs 0 through 900) on the 10 000 CIFAR-10 test images. The per-neuron maximally exciting input (MEI) analysis (§3.1) uses a smaller compute-bounded subset of (sublayer, epoch) cells (final $n = 179$; full sampling grid in Supplementary §S4).

Visual cortex. We used the publicly available longitudinal two-photon calcium imaging dataset from Goltstein et al. [2021], in which 10 mice were imaged across 9 visual cortical areas during multi-week training in a visual category-learning task. Mice were trained to assign drifting gratings, parametrized by orientation and spatial frequency, to one of two information-integration categories. We compared two in-task stages: the first baseline session before training ("baseline 3", $K = 12$ distinct stimuli) and the last session after the categories were learned (learned stage, $K = 10$ distinct stimuli); the category stage comprises all sessions with these $K = 10$ stimuli, before and after the learning criterion of Goltstein et al. [2021]. We analyzed 5 of the 10 mice in the deposited dataset (F02, F03, F04, K01, K02), using all 21 of their mouse-area recordings across 8 areas. At each stage we used all regions of interest of the first imaging plane of the session, without matching neurons across sessions, so each comparison is between populations of the same area, which need not contain the same neurons. We used the deposited preprocessed traces and behavioral metadata.

Hippocampus. 16 mice were imaged on 4 consecutive days during free exploration of a square open-field arena enriched with floor objects and wall cues. Activity in dorsal CA1 (hippocampal field CA1) was imaged with a head-mounted single-photon miniscope (GCaMP6s, 20 or 30 Hz; 26 and 38 sessions) synchronized to overhead video tracking. Calcium traces were extracted with a constrained non-negative matrix factorization for microendoscopic data (CNMF-E) pipeline built on the CaImAn framework [Pnevmatikakis et al., 2016, Pnevmatikakis and Giovannucci, 2017, Zhou et al., 2018]; components were manually curated and matched across the four days with CellReg [Sheintuch et al., 2017]. After matching, each mouse contributed on average 200 matched neurons (median 158, range 29–435); three sessions were excluded for severe freezing, leaving 61 sessions across all 16 mice. Activity and behavior entered the analysis at every fifth frame (4 or 6 Hz). Full protocol in Supplementary §S1.

2.2 Shared analytical scaffold

All three systems were analyzed with one analytical scaffold and substrate-matched estimators (Table 1). The information plane of Shwartz-Ziv and Tishby [2017] has two axes, input retention $I(X;T)$ and target relevance $I(T;Y)$, where T is the observed population activity, X the external input and Y the task-relevant target. For each system, at each time-resolved checkpoint (epoch for SNN, session for cortex and hippocampus), we extracted the population activity vector T and trained an autoencoder on it to obtain a low-dimensional latent $Z = f(T)$ that is used as an estimator probe, as in compression-based estimation [Goldfeld et al., 2019]. Z is a bottleneck in the architectural sense (a dimensionality reduction that compresses T), but it is not the bottleneck variable of the IB objective. We then estimated two paired quantities [Cover and Thomas, 2006]: the relevance $I(Z;Y)$, how much Z carried about the task-relevant target, and a substrate-appropriate input-axis retention quantity ($H_{\text{gauss}}(T)$ for SNN; $R^2(Z \to X_{\text{phys}})$ for cortex with $X_{\text{phys}} = (\sin 2\theta, \cos 2\theta, \log_2 \mathrm{SF})$ encoding grating orientation θ and spatial frequency SF; $I(X_{\text{ego}}; Z)$ for hippocampus with X_{ego} the 6-component egocentric state of speed, angular velocity, and four arena-object distances; full definitions in Supplementary §S1). The autoencoder was trained with a combined loss summing reconstruction mean squared error (MSE) and, weighted by λ, a supervised term (classification cross-entropy for SNN and cortex, MSE on the current (x, y) position for hippocampus); the supervision ($\lambda > 0$) prevents Z from losing target-relevant structure during reconstruction. Because $Z = f(T)$, $X \to T \to Z$ is a Markov chain, and by the data-processing inequality $I(X;Z) \le I(X;T)$ and $I(Z;Y) \le I(T;Y)$: information read from Z is a lower bound on the corresponding coordinate of T, and every probe is limited to the information that a decoder can extract from T. For the SNN, where $H_{\text{gauss}}(T)$ is computed on T before the autoencoder, any compression observed belongs to the network's own representations across training and cannot come from the probe; for the biological systems the probe size and training were

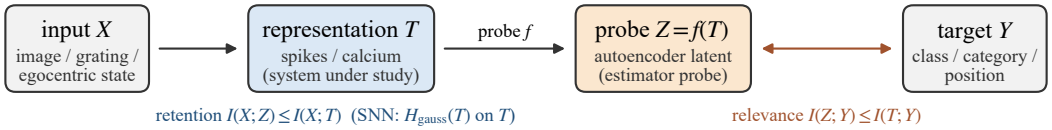

Table 1: Variables and estimators. Top: the observed representation T is read by a separate estimator probe $Z = f(T)$ (bounds in the text). Bottom: variables and the estimator on each axis per system (R^2 is a decoding score, not an information bound). Figures 1, 3 and 4 plot these estimates.

System	Input X	Representation T	Target Y	Retention estimate	Relevance estimate
SNN	CIFAR-10 image	spike counts, 17 sublayers	10 classes	$H_{\text{gauss}}(T)$ on T	KSG $I(Z;Y)$
Cortex	grating (θ, SF)	calcium, 157–308 neurons	binary category	$R^2(Z \to X_{\text{phys}})$	usable information
Hippocampus	egocentric state (6 components)	CA1 calcium, 29–435 neurons	position, 10×10 bins	Gaussian-copula $I(X_{\text{ego}}; Z)$	confusion-matrix MI

held fixed across sessions, and the retention decline persisted without the supervised term ($\lambda = 0$; Supplementary §S9).

The probe dimension d_Z was the only pipeline parameter set by the target and sample size of each substrate rather than tuned: $d_Z = 16$ for SNN (10-class CIFAR target; robust to $d_Z = 8$, see Supplementary §S4), $d_Z = 8$ for cortex (binary category target, populations of 157–308 neurons), $d_Z = 16$ for hippocampus (10×10 spatial target, populations of 29–435 neurons; the median per-mouse slope kept its sign on both axes at $d_Z = 8$ and 32, Supplementary §S9.2). Optimizer, hidden widths, training epochs, and minibatch sizes are listed in the full hyperparameter table in Supplementary §S2.

Primary results used $\lambda = 1$; an unsupervised $\lambda = 0$ control verified the learning signal was not an AE-supervision artifact (Supplementary §S4, §S6, §S9).

2.3 Estimators, cross-validation, and bias correction

Estimator validity. Saxe et al. [2018]'s binning-artifact critique of Shwartz-Ziv and Tishby [2017] was structurally inapplicable to our substrates: SNN spike counts are discrete with deterministic forward pass (so $I(X;T) = H(T)$ is finite in principle), and biological systems are stochastic channels via calcium imaging (cf. Goldfeld et al. [2019]). For SNN we estimated the input-axis quantity via a Gaussian-form differential entropy

$$H_{\text{gauss}}(T) \;=\; \tfrac{1}{2} \sum_{k\,:\,\lambda_k > \tau} \log_2\!\big(2\pi e\, \lambda_k\big), \tag{1}$$

where $\{\lambda_k\}$ are the eigenvalues of the per-layer sample covariance of channel-pooled T (channel-pooled: spike counts summed over timesteps and averaged over spatial positions, so the feature dimension is the channel count, 64–512, well below the 10,000 test images, so the empirical covariance is well-conditioned) and $\tau = 10^{-10}$ excludes eigenvalues at the level of numerical noise. We use $H_{\text{gauss}}(T)$ as a covariance-summary proxy for representation spread along the input axis, not as an exact estimate of the discrete entropy of spike counts; conditioning and dequantization checks are in Supplementary §S4. The other estimators were matched to what each substrate's statistics permit (Table 1). The SNN relevance axis used KSG [Kraskov et al., 2004], which the near-deterministic, large-sample regime supports. In the biological recordings unbiased MI cannot be estimated reliably at these population sizes and trial counts, so both relevance axes used decoder-based usable information [Xu et al., 2020, Kleinman et al., 2025]: temperature-scaled logistic regression for cortex, confusion-matrix MI of a logistic decoder for hippocampus. The hippocampal retention axis used Gaussian-copula MI, chosen over k-nearest-neighbour estimation because of the latter's small-sample bias (Supplementary §S9). The cortical retention axis used the decoding score R^2 (RidgeCV). Robustness checks in Supplementary §S4, §S6, §S9.

Cross-validation. Splits matched each paradigm (SNN, fixed CIFAR-10 train/test split; cortex, stratified 5-fold over trials; hippocampus, contiguous 70/30 temporal split, since random folds would

leak through calcium autocorrelation) and were applied identically to real and shuffled Z (Supplementary §S3.2).

Bias correction. The cortical relevance axis and both hippocampal axes were bias-corrected via per-session shuffle subtraction [Paninski, 2003], $I_{\text{corrected}} = I_{\text{raw}} - I_{\text{shuffle}}$, with paradigm-appropriate shuffles (trial-axis roll for cortex, time-axis roll for hippocampus; details in Supplementary §S3). The SNN relevance axis is the raw estimate, because its label-permutation null sat at the estimator's floor of zero (Supplementary §S3.1).

Significance testing and multiple-comparison policy. Cohort-level claims report Wilcoxon signed-rank tests (one-sided where a directional prediction exists, two-sided for cluster comparisons) with effect sizes; per-mouse claims add a binomial sign-test (Supplementary §S7.2). Per-area tests across cortical areas (Section 3.2) and per-component tests within the 6-dimensional egocentric state (Section 3.3) are exploratory and uncorrected; V1 is an a priori directional prediction (Section 3.2).

3 Results

3.1 Spiking ResNet-18 on CIFAR-10

In the spiking ResNet-18, two regularities emerged: a clean two-phase Tishby trajectory localized to the deep layers, and per-neuron tuning sharpness covaried with population information increasingly tightly with depth.

A two-phase information-plane trajectory localized to deep layers. Figure 1 (top row) shows the information-plane trajectory $I(Z;Y)$ versus $H_{\text{gauss}}(T)$ across training for five sublayers along the network depth. In the early-to-middle layers (L2.1, L3.0, L3.1) the trajectory was nearly monotonically right-and-up: both quantities grew with training, with no compression. In the deep layers (L4.0, L4.1) the within-layer training trajectory turned: $H_{\text{gauss}}(T)$ rose during the first $\sim$70–125 epochs, reached a maximum, and then *reversed leftward* while $I(Z;Y)$ rose or saturated; $H_{\text{gauss}}(T)$ turned down in all four L4 sublayers, after an early rise in three of them. The two-phase Tishby pattern was therefore present, but layer-specific: it characterized the deep part of the network, not training as a whole. In these four sublayers the decodability of six image statistics outside the class target (brightness, contrast, per-channel means and high-frequency energy) slowed, stalled or fell over training, whereas it rose throughout in the other 13; class decodability rose fastest in these four of all 17 sublayers (Supplementary §S4.8).

The same data viewed across layers (Figure 1, bottom row) revealed a temporal asymmetry between geometry and target relevance. Both intrinsic dimension (ID: the estimated dimensionality of the representation manifold, two-nearest-neighbour estimator, Ansuini et al., 2019; the geometry axis complementary to the two information axes) and population $I(Z;Y)$ developed hunchback profiles across depth, tracked via per-epoch parabolic fit (Supplementary §S4). The ID peak stabilized early: it settled in stage 2 by epoch ≈ 30 and moved only from L2.0 to L2.1 thereafter. The $I(Z;Y)$ peak formed later (detectable from epoch 30 onward) and drifted strongly rightward, traveling from L2.1 at epoch 30 to L3.1 by epochs 700–900. The hunchback in ID reproduced Ansuini et al. [2019] for an SNN substrate; the hunchback in $I(Z;Y)$ and its rightward drift are, to our knowledge, novel for classifier-based mutual information at modern scale.

Per-neuron tuning sharpness covaried with population information across depth. We next asked whether the population-level information geometry was reflected at the per-neuron level. For each neuron we used the MEI sharpness score ρ_n of Pospelov et al. [2025] (the relative perplexity of the per-class probability distribution evaluated at the neuron's maximally-exciting input; lower ρ_n indicates a more class-specific MEI) and correlated it, layer by layer, with the class information $I(s_n;Y)$ of the neuron's spike count s_n on test stimuli.

Figure 2 (left) shows that at epoch 700 both per-neuron and population aggregations peaked at L3.1, the same sublayer where population $I(Z;Y)$ peaked in Figure 1. Across all (sublayer, epoch) pairs, mean per-neuron MEI sharpness and population $I(Z;Y)$ were correlated at $r = -0.61$ ($p \ll 10^{-3}$): layers carrying more class information contained neurons whose MEIs were, on average, more class-specific.

The within-layer correlation between $I(s_n;Y)$ and MEI sharpness *strengthened monotonically with depth* from L1.1 on (Figure 2, right). In shallow layers per-neuron MI varied little and the correlation

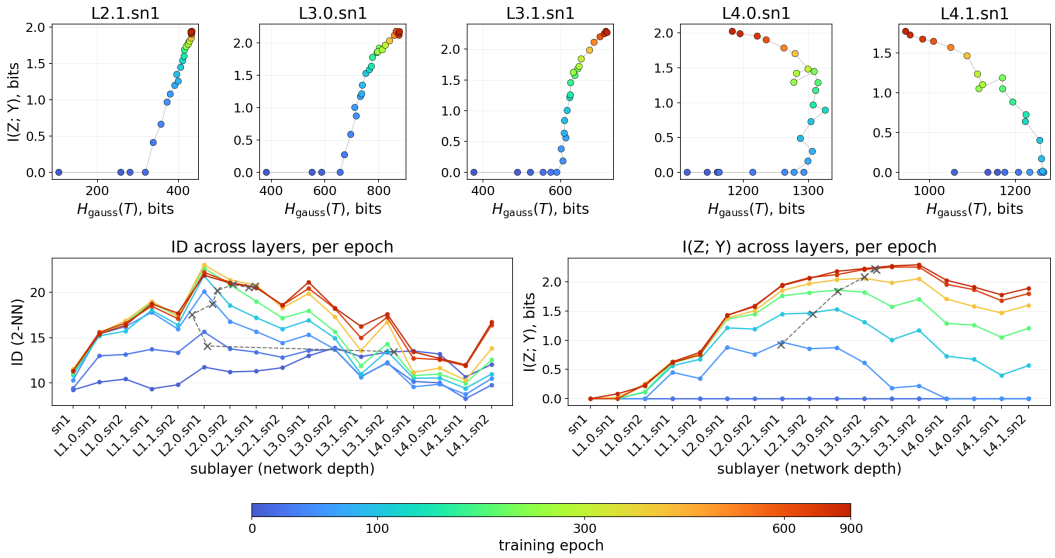

Figure 1: Two-phase compression localized to deep layers (top); the $I(Z;Y)$ peak drifted to L3.1 while the ID peak stayed shallow (bottom). Layer names follow ResNet-18 stage.block: L4.0 and L4.1 are the two residual blocks of stage 4; each block contains two spike layers (sn1, sn2), which with the stem spike layer sn1 gives the 17 analyzed sublayers. **Top row.** Information-plane trajectories $I(Z;Y)$ vs. $H_{\text{gauss}}(T)$ for the first spike layer of five blocks (L2.1.sn1–L4.1.sn1); each point one epoch checkpoint, chronologically connected, colored by training epoch. **Bottom row.** Layer-wise profiles of intrinsic dimension (left) and $I(Z;Y)$ (right) at eight epoch checkpoints (epochs 5 to 900); one line per epoch. Gray crosses and dashed line: per-epoch parabolic-fit peak positions (Supp. §S4).

was weak (sn1: $r = -0.15$); by L3.0 it had tightened to $r = -0.42$, and in the deep layers it was strongly negative (L4.0: $r = -0.66$; L4.1: $r = -0.75$). The depth gradient was the dominant pattern: in deeper layers, the neurons that carried more class information were increasingly the same neurons whose MEIs were more class-specific. The peak-relevance layer L3.1 was the one local exception to this trend, with the within-layer correlation dropping to $r \approx 0$ at epoch 700 (Figure 2 right inset). This layer also hosted a sub-population of broadly-tuned but class-aware neurons larger than in the neighboring layers (Supplementary §S5); we interpret the L3.1 decoupling in §4.

3.2 Mouse visual cortex activity during multi-week category learning

In the multi-week visual category learning paradigm of Goltstein et al. [2021], mice learned to sort drifting gratings, which varied in orientation and spatial frequency, into two categories and reported the category by licking left or right; neurons in each visual area were imaged before training and after the categories were learned (§2.1). The input axis, $R^2(Z \to X_{\text{phys}})$, measures how well the population latent predicts the grating's orientation and spatial frequency; the relevance axis, $I(Z;Y)$, measures how much it carries about the category. Our analysis recovered the published cohort-level effect in information-plane terms and revealed regularities not visible at the neuron-fraction level. Goltstein et al. [2021] clustered cortical areas into a *semantic-leaning, learning-recruited* group {V1, PM, P, POR, LM, LI} and a *feature-leaning, stable* group {AL, RL} (AM not in our cohort); we adopted these labels throughout.

Information-plane translation of the published cohort effect. Across 21 mouse-area pairs, bias-corrected category information $I(Z;Y)$ rose from naive (baseline-3) to trained (last learned session) by a mean of $+0.255$ bits (Wilcoxon $p = 0.0006$, 16/21 positive). Within the semantic-leaning cluster the rise was $+0.266$ bits (13/16 positive, one-sided $p < 0.001$); the feature-leaning cluster rose $+0.22$ bits ($n = 5$ pairs from 5 distinct mice, 3/5 positive, $p > 0.1$). Direction was preserved under the paper-text cluster restriction (Supplementary §S7); a linear mixed-effects model with mouse as random intercept on the 21 pairs gave $p = 1.8 \times 10^{-4}$, confirming the result was not an artifact of within-mouse correlation (Supplementary §S6).

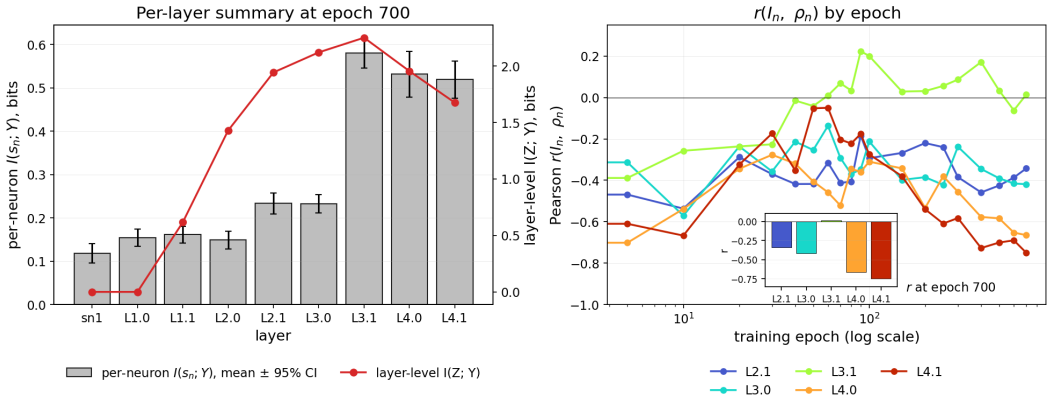

Figure 2: Within-layer coupling between per-neuron MI and MEI sharpness strengthened monotonically with depth, with the peak-relevance layer L3.1 as the local exception. **Left.** Per-layer summary at epoch 700: bars give mean per-neuron $I(s_n; Y) \pm 95\%$ CI (left axis); red line gives population $I(Z; Y)$ (right axis). **Right.** Pearson $r(I_n, \rho_n)$ vs. training epoch (log scale) for five sublayers (L2.1–L4.1, depth-colored turbo). Inset: same correlation at epoch 700. Both panels use the first spike layer (sn1) of each block.

Pipeline artifacts were ruled out by three internal checks (Supplementary §S6). The published per-neuron POR-specific generalized linear model (GLM) rise [Goltstein et al., 2021] is a within-neuron disentanglement outside our population framework, consistent with our small non-significant POR cohort rise ($+0.155$ bits, 2 of 4 pairs positive).

Population-level regularities not visible at the neuron-fraction level.
V1 across all mice. Goltstein et al. [2021] identified V1 as one of two areas with Bonferroni-significant per-area learning effect at the per-neuron level (corrected $P = 0.036$), and assigned V1 to the semantic-leaning, learning-recruited cluster. We therefore took V1 as an a priori directional prediction for the population-information analysis. Every animal in our cohort showed positive learning-related ΔI in V1 (5/5 mice; mean $+0.500$ bits, range $+0.289$ to $+0.706$ bits across mice). With $n = 5$ we report this descriptively, without an independent significance test; per-mouse statistics (one-sided binomial sign-test) appear in Supplementary §S7.2.

Depth-graded information-plane motion. The information-plane trajectories showed graded behavior across the cortical hierarchy (Figure 3, left). The V1 cohort centroid moved up-and-right ($\Delta R^2 = +0.30$, $\Delta I = +0.50$, $n = 5$): both axes grew (consistent with the fitting-phase signature). The semantic-leaning cluster ex-V1 moved predominantly along the relevance axis ($\Delta R^2 \approx +0.01$, $\Delta I \approx +0.16$, $n = 11$). The feature-leaning cluster {AL, RL} drifted up-and-left ($\Delta R^2 = -0.07$, $\Delta I = +0.22$, $n = 5$), the canonical IB compression direction. Per-pair variability was large (3 of 5 pairs $\Delta R^2 < 0$; Wilcoxon $p = 0.62$, two-sided), so cluster direction was not statistically resolved at this cohort size (cluster-boundary checks in Supplementary §S7).

Decoder–behavior gap. The cortical population code carried category information substantially beyond behavioral readout (Figure 3, right). In the category stage ($K = 10$ distinct stimuli), decoder accuracy from cortical Z exceeded behavioral accuracy in all 5 mice (mean gap $+11.5$ pp; one-sided per-mouse Wilcoxon signed-rank on area-averaged decoder accuracy $p = 0.031$, $n = 5$; linear mixed-effects model with mouse as random intercept on the 21 pairs $p = 1.2 \times 10^{-6}$; cluster-aware analysis in Supplementary §S8). The gap held in poorly-trained animals (e.g. F02 V1: 74% decoder vs. 58% behavior, $K = 10$).

3.3 Mouse hippocampus activity during multi-day place learning in open field

CA1 sits as the downstream allocentric stage of a distributed egocentric-to-allocentric circuit [Wang et al., 2018, Alexander et al., 2023]. This transformation maps onto the information plane: $I(X_{\text{ego}}; Z)$ measures input retention (egocentric sensorimotor information preserved in the AE latent Z derived from the matched-across-days CA1 population T), and $I(Z; Y_{\text{alloc}})$ measures relevance to allocentric position binned to a 10×10 arena partition. Input compression discards egocentric inputs

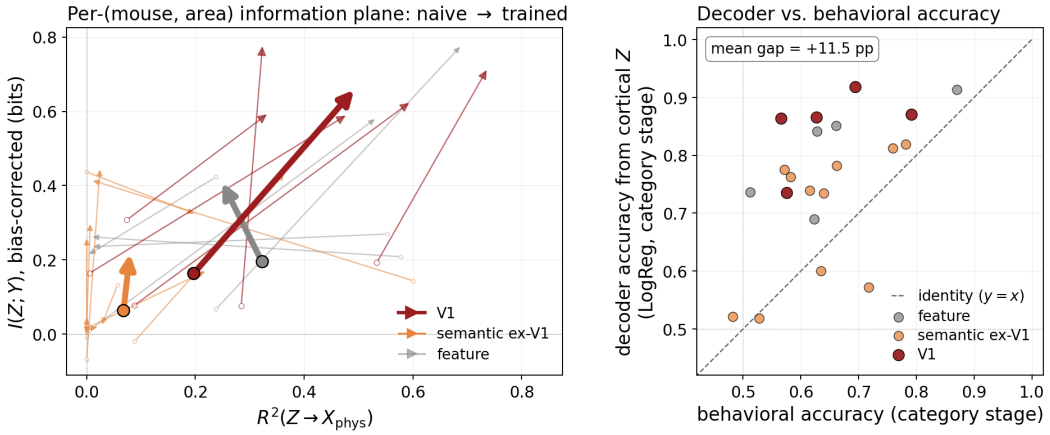

Figure 3: Information-plane trajectories and decoder–behavior gap in mouse visual cortex. **Left.** Per-(mouse, area) trajectory naive→trained on $R^2(Z \to X_{\text{phys}})$ and bias-corrected $I(Z;Y)$. Thin arrows: individual pairs; thick arrows with circles: cluster centroids. Red: V1; orange: semantic-leaning cluster ex-V1 = {PM, P, POR, LM, LI}; gray: feature-leaning cluster = {AL, RL}. **Right.** Decoder accuracy from Z vs. behavioral accuracy in the category stage. Dashed: $y = x$. $n = 21$ mouse-area pairs, 5 mice; cluster assignments follow Goltstein et al. [2021].

not needed for the allocentric readout; output relevance refines the allocentric encoding. Because egocentric distances and allocentric position are geometrically related, we checked that Z carries position information beyond the egocentric state: it did in 59 of 61 sessions (target-conditioned relevance $I(Z; Y_{\text{alloc}} \mid X_{\text{ego}})$, Supplementary §S9; further confounder checks in Supplementary §S11.3–§S11.5).

Bias-corrected information-plane motion across 4 days of familiarization.
Relevance axis: $I(Z;Y_{alloc})$ rose with experience. Across 16 mice over 4 days, hippocampal Z carried more information about allocentric position (Figure 4, left). The cohort-mean trajectory rose from $I(Z;Y_{\text{alloc}}) = 0.54$ bits at day 1 to 0.79 bits at day 4; per-mouse linear slope was positive in 11 of 16 mice (one-sided Wilcoxon signed-rank test, $p = 0.037$; session-inclusion sensitivity in Supplementary §S11). The direction was robust across decoder choice (logreg vs MLP, per-mouse $r = 0.83$) and across the supervised-vs-unsupervised AE regime ($\lambda = 0$ control gave 13/16 rising, Wilcoxon $p = 0.019$). Bias correction was essential, because occupancy concentration with experience depresses I_{shuffle} faster than I_{raw}, leaving $I_{\text{corrected}}$ rising (Supplementary §S10, §S9.5).

Input axis: aggregate compression borderline, structurally localized at component level. The aggregate $I(X_{\text{ego}}; Z)$ compression was borderline: 10/16 mice showed declining slope (Wilcoxon $p = 0.047$); the cohort-mean dropped from 1.24 bits at day 1 to 1.01 bits at day 4. The aggregate p was fragile to session inclusion (sensitivity in Supplementary §S11). The borderline aggregate had a structural explanation: $I(X_{\text{ego}}; Z)$ averages over six components with very different baselines. Per-component analysis (Figure 4, right) showed that compression was confined to some components. All four object distances had negative mean per-mouse slopes, d_{obj2} significantly (13 of 16 mice, one-sided Wilcoxon $p = 0.019$; d_{obj1}, d_{obj3}, d_{obj4} declined in 11, 9 and 10 of 16 mice). Speed and angular velocity showed no significant change because they were not substantially encoded at day 1, and there is nothing to compress when the variable is not present in Z to begin with. A raw-calcium decoder confirmed this was a property of CA1 and not an artifact of the probe: speed, angular velocity, and head direction were not predictable from CA1 activity at the recorded population level (median $R^2 \ll 0$, versus $R^2 > 0$ for the d_{obj1} positive control; Supplementary §S11.3), consistent with their dedicated encoding upstream of hippocampus [Kropff et al., 2015, Taube et al., 1990, Taube, 1995]. None of the six components showed a significantly positive slope, and the compression signal was concentrated in the egocentric subspace that Z encoded most strongly at day 1. A per-session correlation analysis of the latent confirmed what Z encoded: allocentric (x, y) primary, object distances $d_{\text{obj1..4}}$ secondary, speed and ω near zero, robust to λ (Supplementary §S11.4).

Independence from behavioral habituation. As mice familiarized with the arena their spatial occupancy concentrated, lowering the entropy $H(Y_{\text{alloc}})$ of allocentric position, which mechanically de-

presses $I(Z; Y_{\text{alloc}})$ since $I \leq H(Y)$. A linear mixed-effects model with day plus seven behavioral controls including $H(Y_{\text{alloc}})$ showed that the day coefficient grew from $+0.086$ bits/day (day-only, $p = 0.034$) to $+0.139$ bits/day ($p = 0.004$) once $H(Y_{\text{alloc}})$ was held constant, and stayed above the day-only value with six locomotion-related covariates added ($+0.105$ bits/day): behavioral habituation was masking the encoding refinement, not driving it (full model in Supplementary §S11.5).

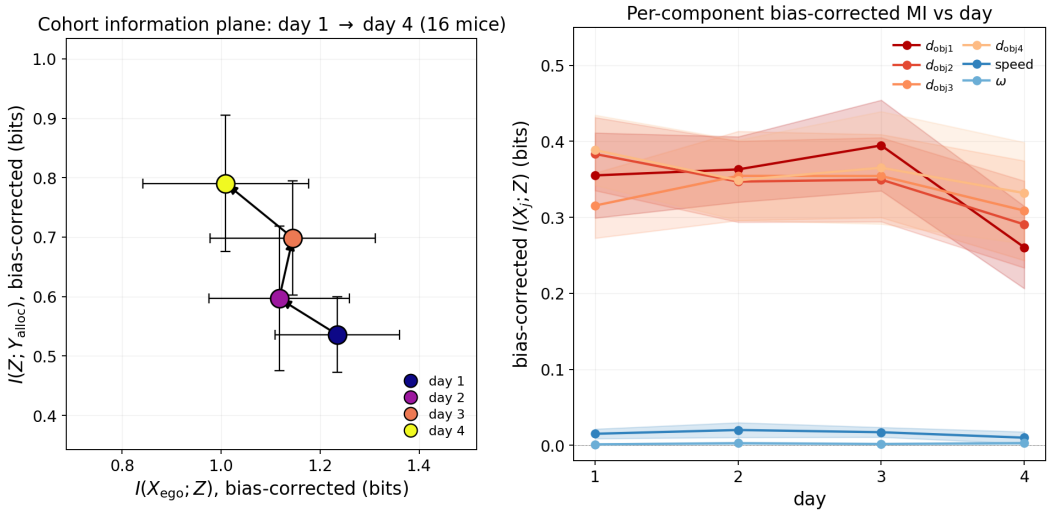

Figure 4: Hippocampal information plane across 4 days of familiarization. **Left.** Cohort-mean information plane (bias-corrected); four day-colored markers (plasma) with SEM bars connected by day-1→day-4 trajectory. Per-mouse direction summary in Supplementary §S9.5. **Right.** Per-component bias-corrected MI $I(X_j; Z)$ (Gaussian copula, as in Supplementary §S11) mean vs day for the six components of X_{ego}, with SEM bands; object distances in warm gradient. 16 mice, 61 sessions.

4 Discussion

We tracked information-plane trajectories during learning in three distinct systems.

In the spiking ResNet-18 (Section 3.1), the trajectory was the cleanest of the three: a two-phase fitting-then-compression sequence localized to the deep layers and was absent in the shallower ones. The ID peak stabilized early at shallower layers while the $I(Z; Y)$ peak drifted rightward across training to settle deeper, suggesting that deeper layers refine task-relevant information atop mid-network geometric representations that had already stabilized. Per-neuron tuning sharpness covaried with population information increasingly tightly with depth, except at the peak-relevance layer L3.1 itself, where the within-layer correlation dropped to $r \approx 0$. The L3.1 exception coincides with two coexisting populations, sharp-tuned and broadly-tuned but class-aware, neither of which shows the coupling within this layer (Supplementary §S5). This complements the IB-as-training-objective work for spiking networks [Buesing and Maass, 2010, Pogodin and Latham, 2020, Yang and Chen, 2023] with the diagnostic counterpart in a deep, gradient-trained SNN, and extends to a spiking substrate the information-plane analyses Butakov et al. [2024] established for ResNet-scale ANNs on CIFAR-10.

In mouse visual cortex (Section 3.2), our analysis recovered Goltstein et al. [2021]'s semantic-cluster cohort effect in information-theoretic terms, validating that population MI reflects real category-coding change [Freedman et al., 2001]. Two regularities emerged beyond the published analysis: V1 gained category information in every animal and moved further in the information plane than either cluster of the other areas, and the cortical code carried category information beyond behavioral readout, a recurring finding in primate sensory neuroscience [Chen et al., 2006, Palmer et al., 2007, Seidemann and Geisler, 2018]. Read together with Goltstein et al. [2021]'s causal demonstration that visual-cortex inactivation abolishes categorization, this gap may suggest that cortex is necessary but not the limiting stage, with the decision bottleneck downstream. Where Kleinman et al. [2025]

document static IB-style organization across cortical areas in already-trained monkeys, our cortex result extends information-plane diagnostics to the learning-time axis.

In mouse hippocampus during arena familiarization (Section 3.3), where the input and relevance axes read as egocentric sensorimotor state and allocentric position, both axes moved with experience: relevance rose and aggregate input information declined, with the aggregate compression borderline at the cohort level but clearly localized at the per-component level in the egocentric-distance subspace encoded by Z. A linear mixed-effects model showed that the relevance-axis rise did not shrink when additional behavioral covariates were controlled. Recent representational-geometric work on longitudinal CA1 has documented complementary reorganizations during learning [Chiossi et al., 2025, Sun et al., 2025]; our information plane provided the information-theoretic decomposition in a less structured, free-exploration paradigm, extending earlier information-theoretic analyses of biological neural data [Meshulam et al., 2019, Palmer et al., 2015].

Two cross-substrate parallels emerged. The most robust cross-substrate observation is that the relevance axis rose with learning in all three systems. Compression along the input axis was resolved where temporal or component resolution permitted: deep layers of the spiking network and the egocentric-distance subspace in hippocampus; in cortex, with only naive-vs-trained endpoint sampling and 5 mice with heterogeneous behavioral acquisition, compression direction was not statistically resolved. Compression is not universal across these systems, which differ in all four together: task, substrate, estimator and dataset. An isolating design, varying one of them at fixed others, is future work; for the SNN it would vary capacity at fixed depth and task, because in our network the sublayers that compress are also the deepest and our design cannot separate compression from depth. We do not extend this to a temporal claim about a universal fitting-then-compression sequence; only the SNN had the dense epoch sampling needed to resolve phase structure.

Why two coordinates. A decline in retention alone does not distinguish shedding input in favour of the target from encoding less of everything. Read jointly with relevance, the deep SNN sublayers, where relevance rose or held while retention reversed, should shed input that is not the target, and the decodability of six image statistics outside both coordinates slowed, stalled or fell there (Section 3.1); V1, in contrast, gained on both coordinates (Figure 3). Across animals the plane was linked to the learning outcome: in visual cortex, mice that learned the task better gained more relevance (Spearman 0.70, 5 mice; Supplementary §S8.4), and in hippocampus, mice with higher day-4 relevance also kept more egocentric input (Spearman 0.70, 14 mice; Supplementary §S9.6).

Three caveats accompany these results. Biological cohorts are modest (5 mice for cortex, 16 for hippocampus) with heterogeneous behavioral acquisition, so per-cluster directional claims should be read as suggestive, not quantitative, and the links between the information plane and the learning outcome of individual animals (Supplementary §S8.4, §S9.6) as outcome-linked and diagnostic, not as a temporal forecast. The three paradigms differ in structure (gradient training, head-fixed discrimination, free behavior) and target entropy: cross-substrate observations should be treated as direction-and-existence claims and read with caution as magnitude comparisons. Finally, dense temporal sampling exists only in the SNN; the fitting-then-compression sequence we resolve there cannot be confirmed or excluded in the biological systems from naive-vs-trained (cortex) or per-day (hippocampus) endpoints, leaving the temporal phase structure of biological information-plane trajectories as the natural target for chronic recording at higher cadence.

The plane also makes testable predictions. Under continual learning we predict a segmented trajectory, one segment per task: within a task it should resemble single-task training; at each task switch, relevance for previously learned targets should step back, so the plane reads out forgetting versus retained competence; and under limited capacity, later tasks should compress earlier-task input structure further.

The information plane offers a substrate-independent coordinate for tracking representation learning. Across a gradient-trained spiking network, a peer-reviewed cortical category-learning paradigm, and a previously unpublished hippocampal recording, the same two axes carried interpretable motion during learning, in the same direction along the relevance axis and with substrate-specific shape and strength along the input axis. The framework's vocabulary of input retention and target relevance translates between gradient-descent dynamics and biological learning at the population level, supplying a shared coordinate connecting two fields long developed independently.

Acknowledgments and Disclosure of Funding

This work was supported by Non-Commercial Foundation for Support of Science and Education "INTELLECT". The authors declare no competing interests.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Andrew S. Alexander, Jennifer C. Robinson, Chantal E. Stern, and Michael E. Hasselmo. Gated transformations from egocentric to allocentric reference frames involving retrosplenial cortex, entorhinal cortex, and hippocampus. *Hippocampus*, 33(5):465–487, 2023. doi: 10.1002/hipo.23513.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pages 531–540, 2018.

Lars Buesing and Wolfgang Maass. A spiking neuron as information bottleneck. *Neural Computation*, 22(8): 1961–1992, 2010. doi: 10.1162/neco.2010.08-09-1084.

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, Alexey Frolov, and Kirill Andreev. Information bottleneck analysis of deep neural networks via lossy compression. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.08013.

Yuzhi Chen, Wilson S Geisler, and Eyal Seidemann. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nature Neuroscience*, 9(11):1412–1420, 2006. doi: 10.1038/nn1792.

Heloisa S. C. Chiossi, Michele Nardin, Gašper Tkačik, and Jozsef Csicsvari. Learning reshapes the hippocampal representation hierarchy. *Proceedings of the National Academy of Sciences*, 122(11):e2417025122, 2025. doi: 10.1073/pnas.2417025122. Preprint: bioRxiv 2024.08.21.608911.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. doi: 10.1002/047174882X.

Robert de Mello Koch and Animik Ghosh. A Two-Phase perspective on deep learning dynamics. *arXiv preprint arXiv:2504.12700*, 2025.

Wei Fang, Yanqi Chen, Jianhao Ding, Zhaofei Yu, Timothée Masquelier, Ding Chen, Liwei Huang, Huihui Zhou, Guoqi Li, and Yonghong Tian. SpikingJelly: an open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40), 2023. doi: 10.1126/sciadv.adi1480.

David J. Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001. doi: 10.1126/science.291.5502.312.

Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308, 2019.

Pieter M. Goltstein, Sandra Reinert, Tobias Bonhoeffer, and Mark Hübener. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nature Neuroscience*, 24(10):1441–1451, 2021. doi: 10.1038/s41593-021-00914-5.

Ishan Kalburge and Máté Lengyel. When sufficiency is insufficient: the functional information bottleneck for identifying probabilistic neural representations. *arXiv preprint arXiv:2512.15671*, 2025.

Michael Kleinman, Tian Wang, Derek Xiao, Ebrahim Feghhi, Kenji Lee, Nicole Carr, Yuke Li, Nima Hadidi, Chandramouli Chandrasekaran, and Jonathan C Kao. The information bottleneck as a principle underlying multi-area cortical representations during decision-making. *eLife*, 2025. doi: 10.7554/eLife.89369.2.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), 2004. doi: 10.1103/PhysRevE.69.066138.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Emilio Kropff, James E. Carmichael, May-Britt Moser, and Edvard I. Moser. Speed cells in the medial entorhinal cortex. *Nature*, 523(7561):419–424, 2015. doi: 10.1038/nature14622.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. doi: 10.1016/S0893-6080(97)00011-7.

Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018. doi: 10.1038/s41593-018-0209-y.

Leenoy Meshulam, Jeffrey L. Gauthier, Carlos D. Brody, David W. Tank, and William Bialek. Coarse graining, fixed points, and scaling in a large population of neurons. *Physical Review Letters*, 123(17), 2019. doi: 10.1103/PhysRevLett.123.178103.

J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971. doi: 10.1016/0006-8993(71)90358-1.

Chris Palmer, Shao-Ying Cheng, and Eyal Seidemann. Linking neuronal and behavioral performance in a Reaction-Time visual detection task. *The Journal of Neuroscience*, 27(30):8122–8137, 2007. doi: 10.1523/JNEUROSCI.1940-07.2007.

Stephanie E. Palmer, Olivier Marre, Michael J. Berry II, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015. doi: 10.1073/pnas.1506855112.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272.

Eftychios A. Pnevmatikakis and Andrea Giovannucci. NoRMCorre: an online algorithm for piecewise rigid motion correction of calcium imaging data. *Journal of Neuroscience Methods*, 291:83–94, 2017. doi: 10.1016/j.jneumeth.2017.07.031.

Eftychios A. Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A. Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha Ahrens, Randy Bruno, Thomas M. Jessell, Darcy S. Peterka, Rafael Yuste, and Liam Paninski. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016. doi: 10.1016/j.neuron.2015.11.037.

Roman Pogodin and Peter E. Latham. Kernelized information bottleneck leads to biologically plausible 3-factor hebbian learning in deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, pages 5171–5180, 2019.

Nikita Pospelov, Andrei Chertkov, Maxim Beketov, Ivan Oseledets, and Konstantin Anokhin. Fast gradient-free activation maximization for neurons in spiking neural networks. *Neurocomputing*, 618:129070, 2025. doi: 10.1016/j.neucom.2024.129070.

Andrew Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Tracey, and David Cox. On the information bottleneck theory of deep learning*. In *International Conference on Learning Representations (ICLR)*, 2018.

Eyal Seidemann and Wilson S. Geisler. Linking V1 activity to behavior. *Annual Review of Vision Science*, 4 (1):287–310, 2018. doi: 10.1146/annurev-vision-102016-061324.

Liron Sheintuch, Alon Rubin, Noa Brande-Eilat, Nitzan Geva, Noa Sadeh, Or Pinchasof, and Yaniv Ziv. Tracking the same neurons across multiple days in ca2+ imaging data. *Cell Reports*, 21(4):1102–1115, 2017. doi: 10.1016/j.celrep.2017.10.013.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Weinan Sun, Johan Winnubst, Maanasa Natrajan, Chongxi Lai, Koichiro Kajikawa, Arco Bast, Michalis Michaelos, Rachel Gattoni, Carsen Stringer, Daniel Flickinger, James E. Fitzgerald, and Nelson Spruston. Learning produces an orthogonalized state machine in the hippocampus. *Nature*, 640(8057):165–175, 2025. doi: 10.1038/s41586-024-08548-w.

Jeffrey S. Taube. Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *The Journal of Neuroscience*, 15(1):70–86, 1995. doi: 10.1523/JNEUROSCI.15-01-00070.1995.

Jeffrey S. Taube, Robert U. Muller, and James B. Ranck, Jr. Head-direction cells recorded from the postsubiculum in freely moving rats. I. description and quantitative analysis. *The Journal of Neuroscience*, 10 (2):420–435, 1990. doi: 10.1523/JNEUROSCI.10-02-00420.1990.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015. doi: 10.1109/ITW.2015.7133169.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control and Computing*, pages 368–377, Monticello, IL, 1999. arXiv:physics/0004057.

Cheng Wang, Xiaojing Chen, Heekyung Lee, Sachin S. Deshmukh, D. Yoganarasimha, Francesco Savelli, and James J. Knierim. Egocentric coding of external items in the lateral entorhinal cortex. *Science*, 362(6417): 945–949, 2018. doi: 10.1126/science.aau4940.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2002.10689.

Shuangming Yang and Badong Chen. SNIB: improving Spike-Based machine learning using nonlinear information bottleneck. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(12):7852–7863, 2023. doi: 10.1109/TSMC.2023.3300318.

Pengcheng Zhou, Shanna L Resendez, Jose Rodriguez-Romaguera, Jessica C Jimenez, Shay Q Neufeld, Andrea Giovannucci, Johannes Friedrich, Eftychios A Pnevmatikakis, Garret D Stuber, Rene Hen, Mazen A Kheirbek, Bernardo L Sabatini, Robert E Kass, and Liam Paninski. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife*, 7, 2018. doi: 10.7554/eLife.28728.

Contents

1 Introduction **1**

2 Methods: a shared information-plane scaffold **2**

2.1 Three systems 2

2.2 Shared analytical scaffold 3

2.3 Estimators, cross-validation, and bias correction 4

3 Results **5**

3.1 Spiking ResNet-18 on CIFAR-10 5

3.2 Mouse visual cortex activity during multi-week category learning 6

3.3 Mouse hippocampus activity during multi-day place learning in open field 7

4 Discussion **9**

S1 Hippocampus experimental and preprocessing details **16**

S1.1 Animals and ethics 16

S1.2 Surgery and viral delivery 16

S1.3 Behavioral procedure 16

S1.4 Behavioral tracking and feature extraction 16

S1.5 Calcium imaging processing 17

S1.6 Session exclusion 17

S2 Autoencoder hyperparameters and training details **17**

S2.1 Per-substrate hyperparameters 17

S2.2 Bottleneck dimension d_Z 18

S2.3 Compute resources 18

S3 Bias-correction shuffles and cross-validation schemes **18**

S3.1 Shuffle nulls 18

S3.2 Cross-validation schemes 19

S4 SNN information-plane robustness **19**

S4.1 Latent-dimension robustness 20

S4.2 Unsupervised regime ($\lambda = 0$) 20

S4.3 Spike-count reconstruction loss 20

S4.4 Per-sublayer peak/end values in deep layers 21

S4.5 Hunchback parabolic-fit method 21

S4.6 Validity of the Gaussian-form estimator: eigenvalue spectrum and dequantization . 21

S4.7 MEI sublayer/epoch coverage 22

S4.8 Auxiliary image statistics: what the reversing sublayers shed 22

S5 SNN per-neuron tuning diagnostics **23**
S5.1 Layer-wise summary at epoch 700 . . . 23
S5.2 Threshold-filtered correlation . . . 24
S5.3 Per-layer (ρ_n, I_n) scatter colored by MEI max-activation . . . 24
S5.4 Interpretation: L3.1 mixes two coexisting coding regimes . . . 24

S6 Cortex information-plane robustness **25**
S6.1 Blocked-CV agreement . . . 25
S6.2 Shift-shuffle null . . . 25
S6.3 Unsupervised regime ($\lambda = 0$) . . . 25
S6.4 Input-axis shuffle null . . . 26
S6.5 Input-axis $R^2(Z \to X_{\text{phys}})$ across training stages . . . 26

S7 Cortex cluster-membership ambiguity and V1 per-mouse statistics **26**
S7.1 Cluster-assignment sensitivity . . . 26
S7.2 V1 per-mouse . . . 26

S8 Cortex decoder-vs-behavior gap with cluster-aware tests **27**
S8.1 Per (mouse, area) data at $K = 10$ . . . 27
S8.2 Per-mouse aggregate . . . 28
S8.3 Cluster-aware test on (mouse, area) pairs . . . 28
S8.4 Learning outcome and relevance gain . . . 28

S9 Hippocampus information-plane robustness **29**
S9.1 Decoder family for the relevance axis . . . 29
S9.2 Alternative probes of the same activity . . . 29
S9.3 Unsupervised regime ($\lambda = 0$): cohort summary on both axes . . . 30
S9.4 Input-axis $\lambda = 0$ vs $\lambda = 1$ contrast . . . 30
S9.5 Per-mouse direction summary . . . 31
S9.6 Better and worse learners . . . 32

S10 Hippocampus bias correction mechanism **32**
S10.1 Coverage shrinkage . . . 32
S10.2 Why raw decoder accuracy gives the wrong sign . . . 33
S10.3 Cohort-mean MI trajectory at $\lambda = 1$ . . . 33

S11 Hippocampus sensitivity analyses and confounder controls **33**
S11.1 Aggregate session-inclusion sensitivity . . . 33
S11.2 Per-component slopes . . . 34
S11.3 Raw-calcium decoder confirms speed/ω/head direction are not in T . . . 35
S11.4 Per-session GLM: what does Z encode? . . . 35
S11.5 Relevance-axis day-coefficient under behavioral-habituation controls . . . 36

S1 Hippocampus experimental and preprocessing details

S1.1 Animals and ethics

We used C57BL/6 mice of both sexes ($n = 16$), aged 3–5 months. Before surgery, animals were group-housed (2–7 per cage) under a 12-hour light/dark cycle with free access to food and water. After surgery, mice were single-housed; other housing conditions were unchanged. All experiments were performed during the light phase of the day (10:00–18:00). All procedures were approved by the host institution's animal-research bioethics commission and followed the NIH Guide for the Care and Use of Laboratory Animals.

S1.2 Surgery and viral delivery

Surgical procedures were carried out in three stages: viral injection, GRIN-lens implantation, and baseplate installation. Anesthesia was induced by intraperitoneal injection of zoletil (40 mg/kg) and xylazine (5 mg/kg). Dexamethasone (4 mg/kg) was administered subcutaneously 5 minutes before surgery to reduce inflammation and prevent cerebral edema. Mice were positioned in a stereotaxic frame, and body temperature was maintained with a heated pad. For viral injection, 300 nL of AAV1.CAG.GCaMP6s was delivered unilaterally into dorsal CA1 (coordinates: AP -1.94 mm, ML $+1.46$ mm, DV -1.2 mm). One week later, a 1-mm-diameter GRIN lens was implanted 200 μm above CA1 and fixed to the skull with cyanoacrylate glue and dental cement; the exposed face was covered with Kwik-Sil. In a third surgery one week later, an aluminum baseplate was placed over the implant site to allow miniscope mounting and fixed with dental cement; a plastic cap was added to protect the lens. Mice were given three weeks to recover before behavioral testing.

S1.3 Behavioral procedure

Two weeks after the final surgery, mice were habituated to handling and to wearing the miniscope: for three consecutive days each animal was placed in its home cage with a connected but non-recording miniscope for 5 minutes per session, allowing it to adapt to the device's weight and presence without additional stress. Following habituation, animals were tested in a square open-field arena ($44 \times 44 \times 44$ cm, gray plastic) for four consecutive days, one 10-minute session per day. The arena was configured as a rich contextual environment, with a semi-transparent silicone floor mat and four floor objects of distinct shape, size, and color; each of the four walls displayed a unique high-contrast visual cue. Throughout each session, the animal's behavior was recorded with a top-mounted video camera while CA1 calcium activity was simultaneously acquired via the miniscope at 20 or 30 Hz (26 and 38 sessions). Behavioral video and miniscope streams were synchronized using shared timestamps, yielding aligned time bases for all downstream analyses.

S1.4 Behavioral tracking and feature extraction

Pose estimation was performed using a DeepLabCut-trained network [Mathis et al., 2018]. For each video, the positions of key body parts (nose tip, head center, body center, tail base, fore- and hind-limb keypoints) were estimated frame-by-frame; coordinates with confidence below 0.95 were excluded and reconstructed by piecewise cubic interpolation from neighboring high-confidence frames. Trajectories were smoothed with a third-order Savitzky–Golay filter (0.25 s window for body center and tail base, 0.1 s window for the remaining body parts).

The information-plane analysis (main paper, Section 2.2) uses six continuous behavioral variables derived from these trajectories: instantaneous speed, angular velocity, and Euclidean distances from the animal's body center to each of the four floor objects ($d_{\text{obj1}}, \ldots, d_{\text{obj4}}$). Each object's position is the mean tracked position inside its zone in that session; an object whose zone is never entered has no position, and its distance is set to a constant zero. Together they form the egocentric sensorimotor state X_{ego}. We use "egocentric" here in the broad sense of self-referenced or non-positional task variables; strictly speaking, the four object distances are scalar (bearing-invariant), and speed and angular velocity are kinematic rather than reference-frame-specific. The relevance-axis target Y_{alloc} is allocentric body-center position (x, y) discretized into a 10×10 grid of square spatial bins covering the arena floor. Additional discrete behavioral states (locomotor speed bin, freezing, rearing, object-zone occupancy) are tracked as covariates and used only in the behavioral-readout analysis of Section S11.5.

S1.5 Calcium imaging processing

Calcium imaging data were processed with a custom analysis pipeline built on the CaImAn framework [Pnevmatikakis et al., 2016]. Each session began with visual inspection to define cropping boundaries. Preprocessing comprised background correction and rigid motion correction with NoRMCorre [Pnevmatikakis and Giovannucci, 2017]. Preprocessed videos were decomposed into spatial footprints and temporal activity traces using constrained non-negative matrix factorization (CNMF-E) [Zhou et al., 2018]. CNMF-E hyperparameters were chosen using predefined quality criteria (spatial-footprint compactness, trace SNR, residual structure) and verified by visual inspection. The main parameter ranges were estimated neuron radius `gSig` 3–6 pixels (at $\sim$1.5 μm/pixel resolution), local-correlation threshold `min-corr` 0.85–0.95, and peak-to-noise ratio `min-pnr` 3–15. Downstream results were qualitatively stable across this parameter range.

After decomposition, putative components were automatically filtered with thresholds `rval-thr` $= 0.95$ and `min-SNR` $= 3$, following the CaImAn guidelines, and all retained components were manually inspected by an experienced annotator to exclude artifacts and duplicates. Only manually verified components were classified as neurons. Fluorescence traces were normalized as dF/F, and the resulting per-session spatial maps were registered across the four imaging days using CellReg [Sheintuch et al., 2017] to track individual neurons longitudinally. After cross-day registration, each mouse contributes on average 200 matched neurons (median 158, range 29–435 across the 16 mice); the same neurons enter all four sessions of a mouse, taken at every fifth frame (4 or 6 Hz).

S1.6 Session exclusion

Three sessions across the 16-mouse $\times$ 4-day cohort showed near-immobility throughout the recording (severe freezing): of the 64 sessions they have the three highest fractions of rest or freezing (87–96% of the session; next highest 78%) and the three lowest distances traveled:

session	distance, arena lengths	cohort percentile	rest or freezing	X_{ego} effective dim
H32 day 4	11.84	1.6%	87%	2
H06 day 4	13.85	3.1%	89%	3
H33 day 3	17.29	4.7%	96%	3

Effective dimensionality counts the components of X_{ego} that vary over the session; the distance to an object whose zone is never entered is constant (above). It reflects unvisited objects as well as immobility, so it does not define the cut: H39 day 4 also has effective dimensionality 3 (18.63 arena lengths, rest or freezing 78%) and is kept, as is H06 day 3 (18.44 arena lengths, dimensionality 5, rest or freezing 64%); Section S11 tests both. All cohort statistics in the main paper use the resulting 61-session set (drop-3 set), retaining all 16 mice.

S2 Autoencoder hyperparameters and training details

The same autoencoder architecture template was used across the three substrates: an encoder $T \to Z$ producing the low-dimensional latent code, a symmetric decoder $Z \to \hat{T}$ trained to minimize reconstruction MSE, and a supervised head $Z \to \hat{Y}$ trained jointly to retain target-relevant structure in the latent. The combined loss is $\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \cdot \mathcal{L}_{\text{predict}}$, with λ controlling the strength of supervision. Primary results used $\lambda = 1$; the unsupervised control ($\lambda = 0$) is reported per substrate within each substrate's information-plane robustness section (Sections S4, S6, S9).

S2.1 Per-substrate hyperparameters

System	encoder hidden	latent d_Z	head hidden	epochs	batch	lr	seed
SNN	$\min(256, 2N)$	16	linear	50	128	10^{-3}	42*
Cortex	128	8	32	80	32	10^{-3}	42
Hippocampus	64	16	32	100	64	10^{-3}	42

The encoder and decoder are MLPs with one hidden layer of the listed width and LeakyReLU non-linearity. The supervised head is a one-hidden-layer MLP (linear $\rightarrow$ non-linearity $\rightarrow$ linear; LeakyReLU in cortex, ReLU in hippocampus) of the listed head-hidden width; for the SNN it is a single linear layer $Z \rightarrow Y$, and the SNN encoder width is $\min(256, 2N)$ for a sublayer with N channels. The head was trained as a classifier with cross-entropy loss for SNN and cortex (categorical targets, $K = 10$ and $K = 2$ classes respectively) and as a regressor with mean-squared-error loss for hippocampus (continuous current-position target (x, y), with the relevance-axis quantity $I(Z; Y_{\text{alloc}})$ estimated post-hoc on the 10×10 binned partition; main paper Section 2.3). The optimizer is Adam in all three pipelines. *For the SNN, seed 42 fixes only the stratified training subsample; autoencoder training itself was not seeded, so a rerun agrees with the reported values up to training noise.

S2.2 Bottleneck dimension d_Z

The probe width d_Z is the one pipeline parameter set by each substrate's target and sample size rather than tuned. We picked d_Z slightly above the target's minimal sufficient dimensionality, leaving slack for class-confidence variation and target-irrelevant residual structure: $d_Z = 16$ for SNN (above the 10 CIFAR-10 classes), $d_Z = 8$ for cortex (binary category, with margin for stimulus residuals), and $d_Z = 16$ for hippocampus (the AE is trained on continuous two-dimensional (x, y), so $d_Z = 16$ has ample slack for positional structure plus task-irrelevant signal; the 100-class binned Y_{alloc} enters only at post-hoc MI estimation, not during AE training). Robustness to alternative d_Z is reported per system in Section S4 (SNN $d_Z = 8$ vs 16) and Section S9.2 (hippocampus $d_Z = 4$, 8, 32); cortex and hippocampus $\lambda = 0$ controls at fixed d_Z are reported in Sections S6 and S9 respectively.

S2.3 Compute resources

All experiments were performed on a single workstation with one consumer GPU (NVIDIA RTX 4090). Per-session AE training and downstream probes completed in seconds for the cortex and hippocampus (each session used a few hundred neurons, and from tens of trials in cortex to a few thousand time samples in hippocampus) and in roughly 1–3 minutes per (sublayer, epoch) cell for the SNN, where the AE was trained on spike-count summaries with input dimension equal to the sublayer's channel count (64 to 512 along the network depth), trained on 10 000 training images and evaluated on the 10 000 CIFAR-10 test images. Compute was dominated by the SNN per-(sublayer, epoch) sweep over 17 sublayers $\times$ 26 epoch checkpoints, repeated across the d_Z robustness check, the λ contrast, and the per-neuron MEI analysis; this set of runs totaled on the order of 100 GPU-hours. The cortex and hippocampus pipelines together added fewer than 10 GPU-hours.

S3 Bias-correction shuffles and cross-validation schemes

This section gives the formal definition of the per-substrate shuffle null used for bias correction (Section S3.1) and the cross-validation scheme used per substrate (Section S3.2). Both choices share an underlying principle: paradigm-matched evaluation that preserves the structural nuisance of the data while isolating the dependency the analysis is testing.

S3.1 Shuffle nulls

Bias correction subtracts a paradigm-matched shuffle null from each mutual-information estimate, $I_{\text{corrected}} = I_{\text{raw}} - \langle I_{\text{shuffle}} \rangle$, so the choice of shuffle determines what the corrected quantity isolates. The general principle is the same across substrates: preserve all nuisance structure that contributes to the bias floor (sample size, temporal autocorrelation, marginal distributions of T and Y) while destroying the dependency between T and Y that the analysis is meant to detect. The implementation is paradigm-specific because the relevant nuisance structure differs across the three datasets.

SNN – random label permutation. For each of five repeats per (sublayer, epoch) cell, the labels Y were randomly permuted across all 10,000 CIFAR-10 test samples while preserving the class marginal, and $I(Z; Y_{\text{shuf}})$ was re-estimated with the same KSG estimator. The permuted estimate sat at the estimator's floor of zero in all 441 cells (one of the 442 cells, layer1.0.sn1 at epoch 0, has no autoencoder because its activity file on the training images is unreadable; unclipped, the estimate was negative in every cell), so subtraction could only raise the curves; the SNN relevance

axis is reported as the raw, conservative estimate. The unclipped null (mean between -8 and -2 bits across cells) is more negative in the deepest sublayers, so subtracting it would move the depth maximum of $I(Z;Y)$ from the middle layers to L4.1.sn1 in 25 of 26 epochs. Because CIFAR-10 test samples are independent and identically distributed, a random label permutation is a sufficient null: it leaves the sampling distribution of T unchanged and breaks only the stimulus–label binding.

Cortex – shift-shuffle. For each of three repeats per session (five for the input axis), and anew in each cross-validation fold, the trial-aligned activity matrix was circularly shifted along the trial axis by a random offset drawn uniformly between 10 trials and $N_{\text{trials}} - 10$ trials. The shift preserved within-trial temporal structure and any across-trial autocorrelation (such as slow baseline drift) while breaking the stimulus–response correspondence. We preferred a circular shift over a random label permutation because permutation would also destroy the across-trial autocorrelation that contributes to the bias floor; subtracting a null with that nuisance structure removed would over-attribute autocorrelation-induced bias to the shuffle term.

Hippocampus – temporal roll. For each of five repeats per session, each neuron's calcium trace was circularly rolled along the time axis by its own random offset, drawn uniformly over the session length, before the train/test split; the autoencoder was then retrained on the rolled activity with the real position target, and both axes were re-estimated on its latent. The roll preserved the temporal autocorrelation of every neuron's calcium trace while destroying the position–activity binding (and the co-activity between neurons) that the analysis is meant to detect. Although the relevance target Y_{alloc} is a discrete 10×10 spatial-bin label, we used a temporal roll on the activity rather than a permutation on Y because Y-permutation does not preserve per-neuron calcium autocorrelation; the resulting non-stationarity bias would be attributed by the corrected MI to genuine encoding rather than to the shuffle null.

S3.2 Cross-validation schemes

Each system used a cross-validation scheme matched to the temporal structure of its paradigm; the same split was applied identically to real and shuffled Z to keep raw and bias-correction estimates aligned.

SNN. Fixed CIFAR-10 train/test split: the autoencoder is trained on a class-stratified subsample of 10,000 of the 50,000 training images (1,000 per class) and all quantities are evaluated on the 10,000 test images. No cross-validation is needed; the single fixed split is the canonical evaluation protocol for CIFAR-10.

Cortex. Stratified 5-fold cross-validation across i.i.d. trials. The Goltstein paradigm is trial-based discrimination, where each trial is an independent stimulus presentation; within-session non-stationarity is therefore not the dominant noise source and a random k-fold split is appropriate. Blocked-CV agreement at $r = 0.97$ per session (chronological folds) confirmed that the 5-fold result was not exploiting trial-to-trial autocorrelation.

Hippocampus. Contiguous 70/30 temporal split (first 70% of each session for training, last 30% for testing). Sessions of free exploration have within-session non-stationarity — the animal's movement repertoire transitions from initial high-variability exploration toward more settled exploration over the course of the ten-minute trial. The temporal split asked whether the population code formed during the exploratory phase generalized to the later, more stereotyped phase. Random or blocked k-fold cross-validation interleaves early and late frames within a session and conflates these behavioral states; a contiguous split keeps them separate.

S4 SNN information-plane robustness

We reported robustness of the SNN information-plane analysis along eight axes: alternative bottleneck dimensions, invariance of the input axis to AE objective and stability of the relevance axis under $\lambda = 0$, per-sublayer numerical detail in the deep layers where the two-phase reversal occurs, the parabolic-fit method used to track hunchback peaks across depth, validity of the Gaussian-form

H_{gauss} estimator (eigenvalue spectrum and dequantization correction), a spike-count reconstruction loss, the (sublayer, epoch) coverage of the per-neuron MEI analysis, and auxiliary image statistics.

S4.1 Latent-dimension robustness

The primary analysis used $d_Z = 16$. We re-ran the full pipeline at $d_Z = 8$ as a sensitivity check on the bottleneck-width choice. Figure S1 shows the per-sublayer information-plane trajectories at $d_Z = 8$, in the same panel layout as the main paper Figure 1 top row. The qualitative pattern was identical: monotonic right-and-up trajectories in L2.1, L3.0, and L3.1, and a clear two-phase reversal of $H_{\text{gauss}}(T)$ in the deep layers L4.0 and L4.1. The two-phase compression result therefore did not depend on the choice of d_Z. The probe size sets when the relevance rise is read out: when H_{gauss} of L4.1.sn1 peaks at epoch 70, its relevance is already 1.46 of a final 2.42 bits at $d_Z = 8$, while at $d_Z = 16$ it is still at the floor of the estimator (0.00 of a final 1.77 bits), so at $d_Z = 16$ the reversal precedes the relevance rise.

S4.2 Unsupervised regime ($\lambda = 0$)

The input-axis quantity $H_{\text{gauss}}(T)$ was computed on the spike-count summary T directly, without any autoencoder, and is therefore invariant to the AE objective by construction. The deep-layer compression — the leftward reversal of $H_{\text{gauss}}(T)$ across late training (main paper Figure 1, top row) — is a property of T alone and does not depend on λ. Only the relevance axis $I(Z;Y)$ passes through the AE and could in principle depend on λ.

We re-trained the autoencoder at $\lambda = 0$ (reconstruction-only objective; classification head removed from the loss) on the five panel sublayers $\{\text{L2.1}, \text{L3.0}, \text{L3.1}, \text{L4.0}, \text{L4.1}\}.\text{sn1}$ across the full 26 epoch checkpoints and computed $I(Z;Y)$ on the unsupervised latent. Per-layer Pearson correlation between $I(Z;Y)$ at $\lambda = 0$ and $\lambda = 1$ across the 26 checkpoints was $r \geq 0.998$ in every panel sublayer (L2.1: $r = 0.998$; L3.0: $r = 0.999$; L3.1: $r = 1.000$; L4.0: $r = 0.999$; L4.1: $r = 0.999$). At the final checkpoint (epoch 900), the relevance values $I(Z;Y)$ at $\lambda = 0$ vs $\lambda = 1$ agreed to within 8.4% (L2.1: 1.77 vs 1.93; L3.0: 2.12 vs 2.18; L3.1: 2.30 vs 2.27; L4.0: 2.01 vs 2.02; L4.1: 1.72 vs 1.77 bits), with L3.1 preserved as the population-information peak in both regimes. The information-plane motion observed in the SNN was therefore not driven by classifier supervision; the unsupervised latent recovered the same trajectory shape and the same relevance peak.

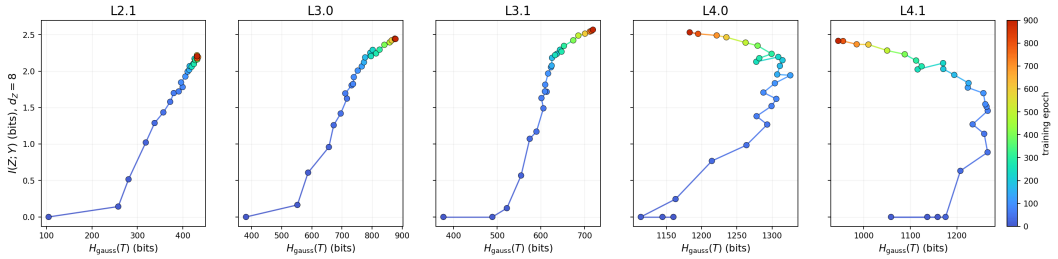

Figure S1: Per-sublayer information-plane trajectories at $d_Z = 8$, mirroring the main paper Figure 1 top row at $d_Z = 16$. Each point is one training-epoch checkpoint, colored by epoch. The input axis $H_{\text{gauss}}(T)$ is computed on T directly and is therefore identical to the $d_Z = 16$ figure; only the relevance axis $I(Z;Y)$ uses the smaller bottleneck. The two-phase reversal in L4.0 and L4.1 (leftward motion of H_{gauss} across late epochs) is preserved.

S4.3 Spike-count reconstruction loss

The autoencoder's reconstruction term is a mean-squared error, which reads spike counts as Gaussian. Only the relevance axis $I(Z;Y)$ passes through the autoencoder, so we re-trained it with a Poisson negative log-likelihood reconstruction term (decoder output read as the log expected spike count), all else unchanged, on the five panel sublayers at epochs 30, 100, 300 and 900, together with a matched mean-squared-error replicate trained in the same run. Under both losses $I(Z;Y)$ rose monotonically over these checkpoints in all five sublayers, and at epoch 900 it peaked across depth at L3.1.sn1 and fell into the L4 sublayers. Across the 20 (sublayer, epoch) cells the two losses

agreed at rank correlation 0.97, and at epoch 900 they differed by at most 0.13 bits. The trajectory and the depth profile of the relevance axis therefore do not depend on the reconstruction model.

S4.4 Per-sublayer peak/end values in deep layers

The four deep-layer sublayers all show a leftward reversal of $H_{\text{gauss}}(T)$ across late training. The peak value, peak epoch, end-of-training value at epoch 900, and the magnitude of the late-training drop are:

sublayer	H_{gauss} peak	epoch at peak	H_{gauss} at epoch 900	drop
L4.0.sn1	1326	125	1183	143
L4.0.sn2	1634	100	1475	158
L4.1.sn1	1266	70	945	321
L4.1.sn2	1230	0	917	313

The compression-phase magnitude grows with depth ($\sim$140 bits in L4.0.sn1 versus $\sim$320 bits in L4.1.sn1).

S4.5 Hunchback parabolic-fit method

To track the position of the across-depth hunchback peak across training, we fit a parabola to the per-epoch profile of each quantity (intrinsic dimension and $I(Z;Y)$) versus sublayer index, and reported the vertex of the parabola as the peak position. The intrinsic-dimension peak settled in stage 2 by epoch ≈ 30 and moved from L2.0 to L2.1 over the rest of training ($+1.8$ sublayers per decade of training epochs, Pearson $r = +0.97$). The $I(Z;Y)$ peak was detectable from epoch 30 onward and drifted strongly rightward across training: $+2.8$ layer-units per decade of training epochs (Pearson $r = +0.98$ between vertex position and $\log_{10}(\text{epoch})$), reaching L3.1 by epochs $\in [700, 900]$.

S4.6 Validity of the Gaussian-form estimator: eigenvalue spectrum and dequantization

The estimator

$$H_{\text{gauss}}(T) = \tfrac{1}{2} \sum_{k:\lambda_k > 10^{-10}} \log_2(2\pi e\, \lambda_k)$$

applies rank-truncation at $\lambda_k > 10^{-10}$ to handle near-singular sample covariance. Two questions of validity arise: whether the empirical covariance is well-conditioned at the sample sizes used, and whether the differential-entropy estimator is meaningful for the discrete-valued spike counts.

Eigenvalue spectrum and rank-conditioning. Channel pooling of the spike-count tensor produced a feature dimension equal to the sublayer's channel count: 64 to 512 across the SNN depth. Each empirical covariance was computed on $n = 10{,}000$ CIFAR-10 test images, so the feature-to-sample ratio stayed at or below about 0.05, well within the well-conditioned regime. At every checked (sublayer, epoch) cell at epoch 900 across all ten dequantization-tested sublayers ($N \in \{128, 256, 512\}$), the truncation threshold $\lambda_k > 10^{-10}$ retained the full eigenvalue spectrum (no eigenvalues fell below 10^{-10}, so $n_{\text{eigs,kept}} = N$ for each sublayer). $H_{\text{gauss}}(T)$ was positive across the entire SNN grid (all 442 cells, 55 to 1634 bits; 104 to 1634 bits in the 260 cells of the dequantization-tested set).

Dequantization correction. At each of the 50 timesteps the recorded value of a channel is the fraction of its P spatial positions that spiked ($P = 256$ in sn1, 64 in stage 1, 16 in stage 2, 4 in stage 3, 1 in stage 4), and T is the sum over timesteps. In stage 4 the values of T are therefore integers in $\{0, 1, \ldots, 50\}$; in earlier stages they are multiples of $1/P$. A Gaussian differential-entropy estimator is not a strict upper bound on the discrete Shannon entropy $H(T)$ without a dequantization adjustment. We checked robustness with the standard uniform-noise dequantization, shifting the covariance by $I/12$ (the variance of $\mathcal{U}[0,1]$ noise on each component) before eigen-decomposition. For the integer values of stage 4 this gives a continuous upper bound on $H(T)$; on the finer grids of earlier stages unit-width noise exceeds the grid spacing, so there the shift is a perturbation test rather than a bound. Per-sublayer comparison across all 26 training epochs, on the ten sublayers spanning the five main-figure depth blocks:

sublayer	N	mean H_{raw}	mean H_{dequant}	mean Δ	relative shift
L2.1.sn1	128	383	388	$+5.4$	$+1.4\%$
L2.1.sn2	128	380	386	$+5.5$	$+1.4\%$
L3.0.sn1	256	749	758	$+9.1$	$+1.2\%$
L3.0.sn2	256	857	862	$+5.0$	$+0.6\%$
L3.1.sn1	256	618	636	$+17.9$	$+2.9\%$
L3.1.sn2	256	616	631	$+15.6$	$+2.5\%$
L4.0.sn1	512	1255	1288	$+33.4$	$+2.7\%$
L4.0.sn2	512	1559	1573	$+13.8$	$+0.9\%$
L4.1.sn1	512	1150	1196	$+46.1$	$+4.0\%$
L4.1.sn2	512	1003	1071	$+68.1$	$+6.8\%$

The mean shift was under 7% per sublayer and followed the expected pattern: deeper layers carried more small eigenvalues for which the additive $1/12$ correction was proportionally larger. Single cells can shift more (19 of 260 at or above 7%, up to 41.5% in L2.1.sn1 at epoch 0).

Two-phase trajectory in L4 is preserved. The leftward reversal of $H_{\text{gauss}}(T)$ in the deep layers — the central qualitative result of the SNN information plane — was robust to the dequantization correction. The peak epoch and direction were unchanged in all four L4 sublayers; the magnitude of the late-training drop shrank by 4–19%:

sublayer	raw peak (epoch)	raw end	raw drop	dequant peak	dequant end	dequant drop
L4.0.sn1	1326 (epoch 125)	1183	-143	1352 (epoch 125)	1237	-116
L4.0.sn2	1634 (epoch 100)	1475	-158	1645 (epoch 100)	1493	-152
L4.1.sn1	1266 (epoch 70)	945	-321	1299 (epoch 70)	1025	-273
L4.1.sn2	1230 (epoch 0)	917	-313	1254 (epoch 0)	997	-257

In the early- and middle-depth blocks (L2.1, L3.0, L3.1), late-training segments were nearly flat in both estimators (raw drops 0–5 bits, dequantized drops within 0.15 bit of raw). The correction therefore did not introduce a phantom two-phase pattern in shallow layers, nor did it remove the existing one in deep layers. We used H_{raw} in the main-text figures and report the dequantized values here as a robustness check.

S4.7 MEI sublayer/epoch coverage

The per-neuron MEI analysis paired MEI sharpness scores ρ_n from Pospelov et al. [2025] with the per-neuron mutual-information values $I(s_n; Y)$ from our pipeline. The MEI computation itself — per-neuron gradient ascent on input space to find the maximally exciting input, and the relative-perplexity scoring of the resulting class distribution — was taken from Pospelov et al. [2025], which sampled 64 neurons per (sublayer, epoch) cell on nine spike sublayers spanning the full network depth (`sn1`, `layer{1..4}.{0,1}.sn1`) at 20 of the 26 epoch checkpoints in our latent-MI grid (omitting epochs $\in \{125, 175, 225, 275, 800, 900\}$). This yielded $9 \times 20 = 180$ candidate (sublayer, epoch) cells; the cell $(\text{layer1.0.sn1}, \text{epoch } 0)$ has no population $I(Z; Y)$ (Section S3.1), leaving $n = 179$ paired observations. Our population-level latent-MI analysis ran independently on all 17 sublayers across 26 checkpoints; the per-neuron analysis inherited their smaller compute-bounded sampling grid (nine sublayers, 20 epochs, 64 neurons per cell).

S4.8 Auxiliary image statistics: what the reversing sublayers shed

In L4.0 and L4.1, $H_{\text{gauss}}(T)$ turns down while $I(Z; Y)$ keeps rising. A joint reading is that these sublayers shed input structure that is not the class. To test it on variables that enter neither axis nor the training objective, we decoded six statistics of each CIFAR-10 test image (brightness, contrast, the three per-channel means and high-frequency energy) from the same channel-pooled spike counts T that $H_{\text{gauss}}(T)$ is computed from, with a ridge regression fitted on 5 000 test images and evaluated on the other 5 000, and took the slope of its held-out R^2 (mean over the six statistics) against $\log_{10}$ epoch. The class was read out the same way, as the held-out R^2 of the one-hot class indicators. In the 13 sublayers where $H_{\text{gauss}}(T)$ does not reverse (its drop from the maximum is at most 22 bits there, against 143–321 bits in L4), decodability of these statistics rose over training (slopes $+0.047$ to $+0.139$, every bootstrap interval above zero). In the four reversing sublayers it rose more slowly

in L4.0.sn1 ($+0.025$), stayed flat in L4.0.sn2 ($+0.001$) and fell in the two deepest, L4.1.sn1 and L4.1.sn2 (-0.053 and -0.032); the two groups do not overlap (Supplementary Figure S2). Measured from each sublayer's own $H_{\text{gauss}}(T)$ maximum onward, the slope stayed positive in L4.0.sn1 and L4.0.sn2 ($+0.056$, $+0.013$) and was negative in L4.1.sn1 and L4.1.sn2 (-0.056, -0.044): the deepest block sheds auxiliary image structure, and L4.0 slows its gain. In the same four sublayers class decodability rose fastest of all 17 (slopes $+0.230$ to $+0.252$), so an analysis that decodes only the class ranks them the best layers. Across depth the class slope rises almost monotonically (Spearman $+0.97$) while the auxiliary slope declines (Spearman -0.39, Pearson -0.56): the two coordinates are not one quantity measured twice.

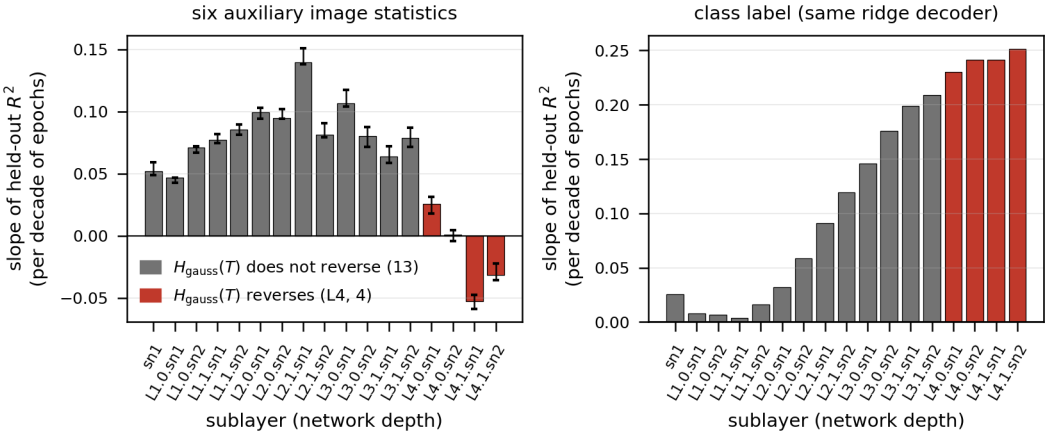

Figure S2: Decodability of six auxiliary image statistics (left) and of the class (right) from channel-pooled spike counts, per sublayer: slope of held-out R^2 (mean over the six statistics, or over the ten one-hot class indicators) against $\log_{10}$ epoch; error bars on the left, bootstrap 95% intervals. Red: the four sublayers in which $H_{\text{gauss}}(T)$ reverses; gray: the other 13.

S5 SNN per-neuron tuning diagnostics

This section reports per-neuron diagnostics for the SNN. Three quantities were paired across neurons:

- $I(s_n; Y)$, the per-neuron mutual information between neuron n's spike count and the class label Y, in bits. The count, a mean over spatial positions and so not an integer before stage 4, was truncated to an integer, and the discrete (plug-in) mutual information with the class label was computed on the 10,000 test images.
- ρ_n, the MEI sharpness from Pospelov et al. [2025]: the relative perplexity of the per-class probability distribution evaluated at the neuron's maximally exciting input. Lower ρ_n indicates a more class-specific MEI. This score and the max-activation below are means over the neuron's three MEI optimization runs.
- MEI max-activation, the activation value (in $[0, 1]$) of neuron n at its maximally exciting input. High max-activation (*sharp-tuned*) means a single input strongly drives the neuron; low max-activation (*broadly tuned*) means no input strongly drives it. We used 0.75 as the split between sharp-tuned and broadly-tuned neurons; this threshold cleanly separated the dominant sharp-tuned population in deep layers (above 0.75: L3.0 81%, L4.0 88%, L4.1 70%) from the broadly-tuned tail concentrated at L3.1 (38% above 0.75).

S5.1 Layer-wise summary at epoch 700

The within-layer Pearson correlation $r(I_n, \rho_n)$, alongside the mean per-neuron mutual information and the broadly-tuned fraction, at training epoch 700 (64 neurons per layer):

layer	$r_{\text{Pearson}}(I_n, \rho_n)$	mean $I(s_n; Y)$, bits	broadly tuned, %
sn1	-0.15	0.12	70.3%
L1.0	$+0.02$	0.15	32.8%
L1.1	-0.19	0.16	93.8%
L2.0	-0.25	0.15	78.1%
L2.1	-0.34	0.23	60.9%
L3.0	-0.42	0.23	18.8%
L3.1	$\mathbf{+0.01}$	$\mathbf{0.58}$	$\mathbf{62.5}\%$
L4.0	-0.66	0.53	12.5%
L4.1	-0.75	0.52	29.7%

Three patterns were visible in the table. (i) The Pearson correlation deepened monotonically from L1.1 (-0.19) to L4.1 (-0.75), with L3.1 ($+0.01$) as the one local exception. (ii) Mean per-neuron MI jumped at L3.1 from 0.23 bits (upstream layers) to 0.58 bits (a $2.5\times$ increase concentrated at this layer); the deeper layers L4.0 and L4.1 retained a similar level (~ 0.52 bits). (iii) The broadly-tuned fraction shrank from 33–94% in upstream layers (sn1–L2.1) to 12–30% in the deep layers (L3.0, L4.0, L4.1), with L3.1 alone retaining a large broadly-tuned fraction (62.5%) resembling upstream layers rather than the deeper layers.

S5.2 Threshold-filtered correlation

To check whether the L3.1 collapse came from the broadly-tuned subset, we restricted each layer to neurons with MEI max-activation ≥ 0.85 (the most sharply tuned subset) and recomputed the Pearson correlation:

layer	r (full population)	r (max-act ≥ 0.85)	n at threshold
L2.1	-0.34	-0.30	8
L3.0	-0.42	-0.55	33
L3.1	$\mathbf{+0.01}$	$\mathbf{-0.21}$	14
L4.0	-0.66	-0.25	43
L4.1	-0.75	-0.52	32

On this subset L3.1's Pearson correlation was -0.21 ($n = 14$, not significant), within the range of the other layers (-0.21 to -0.55), with no order across depth. At this threshold L3.1 is therefore not distinguishable from the other layers, although the subset is small.

S5.3 Per-layer (ρ_n, I_n) scatter colored by MEI max-activation

Figure S3 shows the per-neuron (ρ_n, I_n) distribution at L3.0, L3.1, and L4.0 at epoch 700, with point color encoding MEI max-activation (blue = broadly tuned, red = sharp tuned) and an OLS fit on the full population overlaid in each panel. At L3.0 and L4.0 the population was dominated by sharp-tuned neurons and the fit sloped negative ($r = -0.42$ and $r = -0.66$); at L3.1 the broadly-tuned sub-population is large and the fit is flat ($r = +0.01$).

S5.4 Interpretation: L3.1 mixes two coexisting coding regimes

The depth gradient in $r(I_n, \rho_n)$ has a simple reading. Along the network, neurons increasingly align their feature-detector role with the class structure of the task: a class-specific MEI (low ρ_n) tends to come with high per-neuron mutual information about the class (high $I(s_n; Y)$). The negative correlation deepens with depth as the population commits to class-specific tuning.

L3.1 is the sublayer where the population $I(Z; Y)$ peaked, immediately upstream of the sublayers where $H_{\text{gauss}}(T)$ reverses (main paper Section 3.1, Figure 1). Two facts made L3.1 a transitional layer at the per-neuron level. First, class-information content per neuron jumped at L3.1: mean $I(s_n; Y)$ rose from 0.23 bits at L3.0 to 0.58 bits at L3.1, a $2.5\times$ increase concentrated at this layer. Class encoding emerged here. Second, the broadly-tuned subset persisted from upstream layers — 62.5% of L3.1 neurons had MEI max-activation < 0.75, comparable to L2.1 (60.9%) but markedly larger than at the deeper layers (12–30% in L3.0, L4.0, L4.1) where the population committed to sharp tuning.

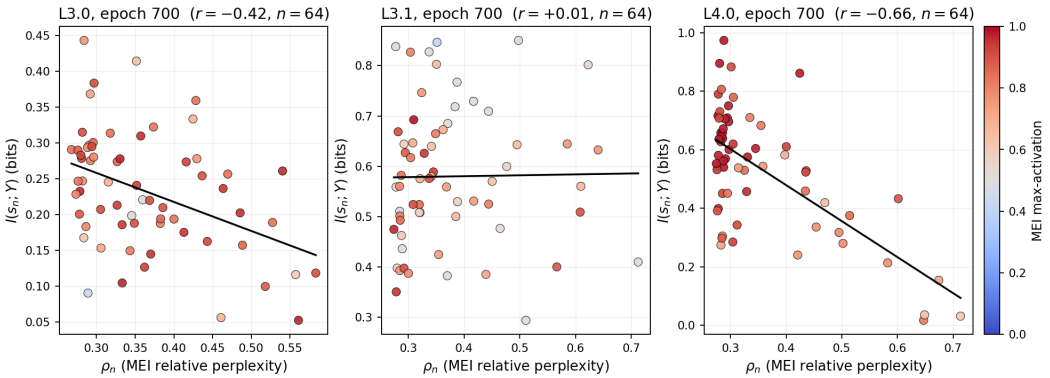

Figure S3: Per-neuron (ρ_n, I_n) scatter at L3.0, L3.1, and L4.0 at training epoch 700 (64 neurons per panel). Point color encodes MEI max-activation: blue = low (broadly tuned), red = high (sharp tuned). Black line: OLS fit on the full population, with the corresponding Pearson r printed in each panel. At L3.0 and L4.0 the population is dominated by sharp-tuned neurons (red) and the fit slopes negative; at L3.1 a large pale (broadly tuned) sub-population is present and the fit is flat ($r \approx 0$).

The combination was most pronounced at L3.1: a substantial population of broadly-tuned neurons that were simultaneously class-aware (40 of 64 neurons, mean $I(s_n; Y) = 0.60$ bits; in L4.1, 19 neurons at 0.45 bits). Within L3.1 the correlation was flat in both groups ($r = -0.03$, $n = 40$, broadly tuned; $r = -0.02$, $n = 24$, sharp-tuned), and on the most sharply tuned subset it was -0.21 ($n = 14$, not significant; Section S5.2). L3.1 thus differs from its neighbors in hosting two populations with similarly high class information, neither of them coupled to MEI sharpness (Figure S3, middle panel).

S6 Cortex information-plane robustness

We reported five robustness checks on the cortex information-plane learning effect (cohort $\Delta I(Z;Y) = +0.255$ bits at $\lambda = 1$, main paper Section 3.2): blocked vs random cross-validation, the shift-shuffle null, the unsupervised $\lambda = 0$ regime, the input-axis shuffle null, and the cohort-level training-stage trend on $R^2(Z \to X_{\text{phys}})$.

S6.1 Blocked-CV agreement

When cross-validation folds were constructed chronologically (blocked) instead of randomly across trials, the per-session $I(Z;Y)$ values agreed with the random-fold values at Pearson $r = 0.97$ (90 sessions), and the per-pair learning changes $\Delta I(Z;Y)$ at $r = 0.94$ (21 pairs). The random k-fold result was therefore not driven by trial-to-trial autocorrelation that random folding would exploit.

S6.2 Shift-shuffle null

Repeating the entire pipeline on a shifted activity matrix (circular shift along the trial axis, breaking stimulus–response binding while preserving trial-structure autocorrelation; Section S3) yielded a cohort-level $\Delta I_{\text{shuffle}} = -0.022$ bits across the 21 mouse-area pairs: slightly negative (17 of 21 pairs; two-sided Wilcoxon $p = 0.0007$) and an order of magnitude smaller than the $+0.255$-bit rise in the real data. The shuffle null did not rise with training, so the bias-corrected cohort statistic reflected a stimulus–response-binding signal, not a trial-structure artifact.

S6.3 Unsupervised regime ($\lambda = 0$)

At the primary regime, the cohort relevance gain was $\Delta I(Z;Y) = +0.255$ bits, positive in all five mice (per-mouse means $+0.024$ to $+0.515$ bits) and in 16 of 21 mouse-area pairs; a linear mixed-effects model with mouse as random intercept on the 21 pairs (maximum-likelihood fit) gave $p = 1.8 \times 10^{-4}$. Removing supervision (bias-corrected with the $\lambda = 1$ shuffle null) dropped the

magnitude to $\Delta I = +0.184$ bits but preserved the direction across animals (4 of 5 mice positive; 16 of 21 pairs; LMM $p = 8.4 \times 10^{-4}$). Supervision sharpened the magnitude rather than creating the signal.

S6.4 Input-axis shuffle null

The cortex input-axis shuffle null was essentially zero: after a circular shift of the activity along the trial axis, regression of the retrained Z onto the stimulus features yielded $R^2 \approx 0$ at every session (per-session R^2_{shuf} median $= 0.000$, range $[0.000, 0.009]$ across the 69 sessions with $K = 10$ or 12), and the cohort-level naive-to-trained change in shuffled R^2 was $\Delta R^2_{\text{shuf}} = +0.0011$ (s.d. 0.0019 across 21 mouse-area pairs). The input axis of main paper Figure 3 is therefore plotted without subtraction.

S6.5 Input-axis $R^2(Z \to X_{\text{phys}})$ across training stages

This check asks how the input axis of Figure 3 moved pair by pair. Between the same endpoints as Figure 3 (first baseline session, last learned-stage session), $R^2(Z \to X_{\text{phys}})$ rose by $+0.060$ on average across the 21 mouse-area pairs (median $+0.025$; 13 pairs rose, 5 fell, 3 stayed at zero; two-sided Wilcoxon $p = 0.35$). The rise was carried by V1 ($+0.30$, 5 pairs); the semantic-leaning areas outside V1 stayed flat ($+0.01$, 11 pairs) and the feature-leaning areas fell (-0.07, 5 pairs, 3 of them falling). Re-expressed as Gaussian mutual information, $-\frac{1}{2}\log_2(1 - R^2)$, which is a transformation of R^2 rather than an independent estimate, the mean per-pair change was $+0.35$ bits in V1, 0.00 bits in the semantic-leaning areas outside V1 and -0.04 bits in the feature-leaning areas. Held-out R^2 was computed per cross-validation fold and floored at zero before averaging over folds, so a fold that predicts X_{phys} worse than its mean counts as carrying no linearly decodable stimulus information.

S7 Cortex cluster-membership ambiguity and V1 per-mouse statistics

This section reports two checks on the cortex result: whether the cluster-2 cohort effect survived a small ambiguity in cluster assignments for two ventral-stream areas, and what the per-mouse breakdown of V1 looked like.

S7.1 Cluster-assignment sensitivity

Goltstein et al. [2021] Fig. 4f stream-test (chi-squared $= 8.58$, $p = 0.0034$) maps ventral-stream areas {LM, LI, P, POR} to cluster 2 and dorsal-stream areas {AL, RL, AM} to cluster 1; we adopted this stream-based assignment (EXT-STREAM, $n = 16$) as primary in the main paper. As a robustness check we report cohort statistics under both EXT-STREAM and the strict text-based assignment STRICT ($n = 13$, V1+PM+P+POR only). C1 = {AL, RL} in both (AM was not in our cohort).

Definition	Cluster (areas in cohort)	n_{pairs}	mean ΔI (bits)	n_{pos}	two-sided Wilcoxon p
STRICT	C2 = {V1, PM, P, POR}	13	$+0.303$	11/13	$\mathbf{0.002}$
	C1 = {AL, RL}	5	$+0.219$	3/5	0.44
EXT-STREAM	C2 = {V1, PM, P, POR, LM, LI}	16	$+0.266$	13/16	$\mathbf{0.001}$
	C1 = {AL, RL}	5	$+0.219$	3/5	0.44

Direction was preserved under both definitions: cluster 2 was significant ($p \leq 0.003$), cluster 1 was non-significant ($p = 0.44$). The stream-based extension added three additional pairs to the semantic-leaning cohort (LM and LI from F03, F04, K02) and the cluster-2 signal remained strong. A direct C2 vs C1 contrast (Mann-Whitney one-sided) did not reach significance under either definition (STRICT $p = 0.42$, EXT-STREAM $p = 0.45$): the magnitude difference was 0.05–0.08 bits but $n_{C1} = 5$ left the contrast underpowered.

S7.2 V1 per-mouse

The V1 result rested on $n = 5$, so we reported the per-mouse ΔI values, leave-one-out stability, and the formal directional test.

mouse	ΔI (bits)	sign
F02	$+0.289$	$+$
F03	$+0.706$	$+$
F04	$+0.525$	$+$
K01	$+0.549$	$+$
K02	$+0.429$	$+$

Cohort statistics: mean $+0.500$, s.d. 0.154, range $+0.289$ to $+0.706$ bits; all five animals positive. Excluding any single V1 mouse left the remaining $4/4$ pairs positive and the cohort-mean ΔI in the range $+0.45$ to $+0.55$ bits across the five LOO subsets. The directional test is one-sided Wilcoxon signed-rank against zero (one-sided per the significance policy stated in main paper Methods Section 2.3): $p = 1/32 = 0.031$, the structural minimum attainable for $n = 5$.

A one-sided binomial sign-test on the same data, 5/5 positive directional outcomes, gave $p = (1/2)^5 = 0.031$, numerically identical to the Wilcoxon floor and independent of any effect-magnitude assumptions. The main-text claim is the descriptive 5/5 directional outcome; the test is reported here for transparency.

S8 Cortex decoder-vs-behavior gap with cluster-aware tests

Behavioral accuracy is a per-animal metric, so we report the decoder–behavior gap at three replication-unit levels: per (mouse, area) pair, per-mouse aggregate, and a cluster-aware test on the per-pair data.

S8.1 Per (mouse, area) data at $K = 10$

In the category stage ($K = 10$ distinct stimuli, sessions before and after the learning criterion of Goltstein et al. [2021]), decoder accuracy ($Z \to$ category, logistic regression on the AE latent) is compared to the animal's behavioral accuracy on the same sessions. Multiple category-stage sessions per (mouse, area) are averaged to one row per pair. The result is 21 mouse-area pairs across 5 mice:

mouse	area	decoder acc	behavioral acc	n_{sess}	gap (pp)
F02	AL	0.737	0.513	2	$+22.4$
F02	P	0.519	0.528	2	-0.9
F02	POR	0.522	0.482	2	$+4.0$
F02	V1	0.736	0.575	2	$+16.1$
F03	AL	0.690	0.623	2	$+6.7$
F03	LM	0.775	0.572	2	$+20.3$
F03	PM	0.739	0.615	3	$+12.4$
F03	POR	0.762	0.582	3	$+18.0$
F03	V1	0.864	0.565	2	$+29.8$
F04	AL	0.851	0.661	1	$+19.0$
F04	LM	0.782	0.662	1	$+12.0$
F04	PM	0.812	0.759	1	$+5.3$
F04	POR	0.572	0.717	1	-14.5
F04	V1	0.918	0.694	1	$+22.5$
K01	POR	0.820	0.781	1	$+3.8$
K01	RL	0.913	0.870	1	$+4.3$
K01	V1	0.871	0.791	1	$+8.0$
K02	LI	0.734	0.639	1	$+9.5$
K02	P	0.600	0.634	1	-3.4
K02	RL	0.841	0.628	1	$+21.4$
K02	V1	0.866	0.627	1	$+23.9$

Cohort mean gap $+11.5$ pp; positive in 18 of 21 pairs; range -14.5 to $+29.8$ pp.

S8.2 Per-mouse aggregate

Aggregating decoder accuracy across areas within each mouse gave one row per independent animal, the natural replication unit for behavioral accuracy:

mouse	n_{areas}	mean decoder acc	behavioral acc	gap (pp)
F02	4	0.628	0.524	$+10.4$
F03	5	0.766	0.592	$+17.5$
F04	5	0.787	0.699	$+8.9$
K01	3	0.868	0.814	$+5.4$
K02	4	0.760	0.632	$+12.8$

All five mice showed a positive gap (cohort mean $+11.0$ pp); one-sided Wilcoxon signed-rank against zero gave $p = 1/32 = 0.031$, the structural minimum attainable for $n = 5$. Excluding F02 (the lowest-behavior animal, $\approx 52\%$ task accuracy) left $n = 4$ mice all positive (mean $+11.1$ pp); the per-mouse outcome was not driven by any single animal.

S8.3 Cluster-aware test on (mouse, area) pairs

We fit a linear mixed-effects model on the 21 pairs with the per-pair gap as outcome, an overall intercept (the cohort-mean gap), and a per-mouse random intercept (allowing each mouse its own baseline level). The intercept estimate was $+11.5$ pp, $p = 1.2 \times 10^{-6}$. The fit returned a between-mouse variance estimate of essentially zero (group variance $= 0.000$, intra-class correlation ≈ 0): the gap was consistent across mice (per-mouse means span $+5.4$ to $+17.5$ pp) but variable across areas within a mouse (e.g. F04 spanned -14.5 pp at POR to $+22.5$ pp at V1). When within-mouse spread dominates between-mouse variance, cluster correction does not shrink the effective sample size, so the cluster-aware p-value matches a naive per-pair z-test; a per-pair t-test gives $p = 9 \times 10^{-5}$.

S8.4 Learning outcome and relevance gain

Visual cortex is the one system in this study with a behavioral measure independent of the neural data, so it allows the comparison of better and worse learners. For each mouse we took the relevance gain (bias-corrected $\Delta I(Z;Y)$ averaged over its recorded areas, Section S6) and its behavioral accuracy in the category stage (Table above). Mice that learned the task better gained more relevance: Spearman 0.70 across the five mice (Supplementary Figure S4); K01 combined the highest accuracy (81%) with the largest gain ($+0.52$ bits), F02 the lowest accuracy (52%) with the smallest ($+0.02$ bits). At $n = 5$ the relation is directional. It links the change in the representation to the learning outcome reached; it is not a forecast of the outcome from early sessions.

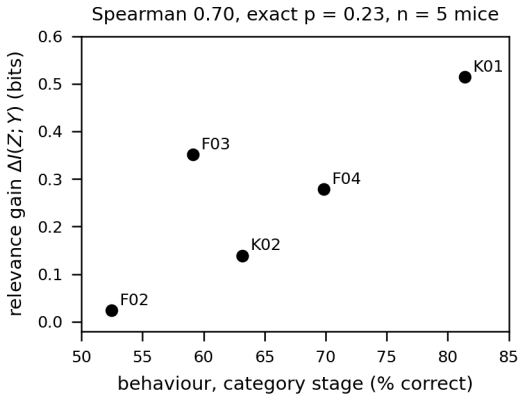

Figure S4: Relevance gain against behavioral accuracy, one point per mouse. Horizontal axis: behavioral accuracy in the category stage, mean over the mouse's recorded areas. Vertical axis: bias-corrected relevance gain $\Delta I(Z;Y)$ from baseline to the learned stage, mean over the mouse's areas.

S9 Hippocampus information-plane robustness

Checks on the hippocampal information-plane analysis, beginning with stability of the relevance-axis estimator under a different decoder family, an unsupervised-regime cohort summary covering both axes, and an input-axis-specific contrast between the supervised and unsupervised regimes.

S9.1 Decoder family for the relevance axis

Does the hippocampal relevance reading depend on the decoder that reads it? The reading has two steps: a logistic regression predicts the 10×10 position bin Y_{alloc} from Z on held-out frames, and mutual information is computed from the confusion matrix between predicted and true bins. We use the confusion matrix because it needs no density estimate of the 16-dimensional Z; the quantity it estimates is the information recoverable by this decoder family, a lower bound on $I(Z; Y_{\text{alloc}})$, and shuffle subtraction removes the finite-sample bias of the estimate. As a control we replaced only the decoder, with a small MLP (one hidden layer of 32 units, weight decay $\alpha = 0.01$, early stopping), on the same saved Z. The rise over days at $\lambda = 1$ gave the same cohort test (11 of 16 mice rising; one-sided Wilcoxon $p = 0.037$), and the per-mouse slopes from the two decoders correlated at Pearson $r = +0.83$. Raw MI was higher with the MLP, which has more capacity (mean 2.24 against 1.91 bits), but the shuffle estimate rose by about the same amount (1.61 against 1.26 bits), so the corrected mean barely moved (0.63 against 0.65 bits), although four mice changed slope sign, two in each direction. In these data a more capable decoder therefore read no more learning-related signal, which points to a limit in what the encoder keeps in Z rather than in the decoder. The retention axis uses the Gaussian-copula estimator because a k-nearest-neighbour estimator is not usable at these sample sizes: on the 16 day-1 sessions KSG ($k = 5$) returned 0.007 bits on average for $I(X_{\text{ego}}; Z)$ (6 + 16 dimensions, 711–1,072 test frames), where the Gaussian copula returned 2.30 bits (0.004 bits after permuting Z), and its local non-uniformity correction, which engages only for k above the joint dimension, gave tens of bits on permuted data as well.

S9.2 Alternative probes of the same activity

We re-estimated both axes with probes that involve no trained network, on the same sessions, splits, five shuffles and bias correction: the first $k = 4, 8, 16, 30$ principal components of T (PCA, linear) and a 16-dimensional Isomap embedding [Tenenbaum et al., 2000] (nonlinear), each fitted on the training fold, real and shuffled data alike (Supplementary Fig. S5). Per-session retention values agreed across the autoencoder ($\lambda = 1$ and $\lambda = 0$), PCA-16 and Isomap-16 (pairwise Spearman 0.84–0.94, 61 sessions), and both axes agreed across the PCA widths (0.78–0.95); relevance values agreed more loosely (0.52–0.91, Isomap being the outlier). At matched size the supervised autoencoder read about as much retention as unsupervised PCA (1.13 and 1.10 bits), and retention grew with the number of components (0.77, 0.98, 1.10, 1.16 bits for $k = 4, 8, 16, 30$), so magnitudes track probe size rather than supervision. The day effect had the same direction across the probe families of this panel: the median per-mouse relevance slope was positive under all seven probes and the median retention slope negative under six of seven. Its strength depended on the probe: the autoencoders of the main size read it most clearly (relevance rising in $11/16$ and $13/16$ mice at $\lambda = 1$ and 0, one-sided Wilcoxon $p = 0.037$ and 0.019), the wider PCA probes more weakly ($11/16$, $p = 0.080$ at $k = 16$; $13/16$, $p = 0.052$ at $k = 30$), and the Isomap embedding, which agreed on retention values, left relevance flat ($8/16$). Since all probes read the same activity T, none can recover information that no decoder captures.

Trained-autoencoder size and a variational probe. We retrained the autoencoder of the main analysis ($\lambda = 1$) with latents of $d_Z = 4$, 8 and 32, and at $d_Z = 16$ a variational autoencoder (VAE) with the same layers and position head plus a Kullback–Leibler term of weight $2/N$ for N neurons, which is the plain VAE objective on this reconstruction scale; its axes were read on the posterior mean, with the variance of the posterior mean above 0.01 in 14.8 of 16 latent dimensions on average. All used the same 61 sessions, splits, five shuffles, estimators and bias correction. The median per-mouse slope had the same sign as for the autoencoder of the main analysis on both axes at $d_Z = 8$, at $d_Z = 32$ and for the VAE (table; Spearman over the 61 per-session values). At $d_Z = 4$ retention kept its direction but the rise in relevance disappeared (7 of 16 mice, median slope -0.021), as it nearly did for PCA at $k = 4$ (8 of 16).

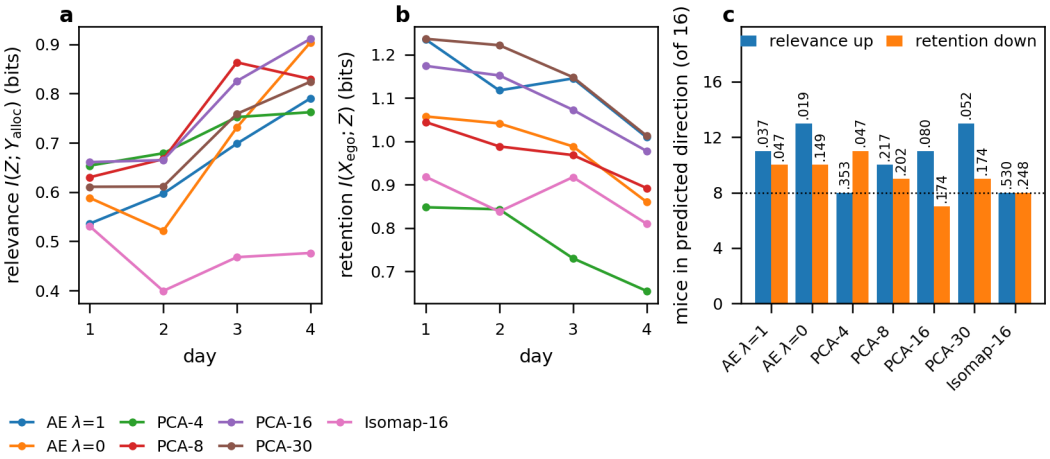

Figure S5: Alternative probes, hippocampus (61 sessions). **a.** Relevance $I(Z; Y_{\text{alloc}})$ per day, mean over mice, for the autoencoder ($\lambda = 1$, $\lambda = 0$), PCA with $k = 4, 8, 16, 30$ components and Isomap-16. **b.** Retention $I(X_{\text{ego}}; Z)$ for the same probes. **c.** Number of mice (of 16) whose per-mouse slope goes in the predicted direction (relevance up, retention down); one-sided Wilcoxon p above the bars; dotted line, half of the mice.

probe	relevance rising mice	relevance rising one-sided p	retention falling mice	retention falling one-sided p	Spearman with $d_Z = 16$ relevance	Spearman with $d_Z = 16$ retention
AE, $d_Z = 16$ (main)	11/16	0.037	10/16	0.047	–	–
AE, $d_Z = 4$	7/16	0.736	9/16	0.106	0.55	0.86
AE, $d_Z = 8$	11/16	0.065	8/16	0.298	0.76	0.88
AE, $d_Z = 32$	12/16	0.019	12/16	0.047	0.84	0.90
VAE, $d_Z = 16$	10/16	0.126	10/16	0.126	0.80	0.89

S9.3 Unsupervised regime ($\lambda = 0$): cohort summary on both axes

We re-ran the full pipeline at $\lambda = 0$ (no supervised head, AE reconstructs T alone). Across the 16-mouse drop-3 cohort, both axes preserved direction:

metric	$\lambda = 0$ (recon-only)	$\lambda = 1$ (Y-supervised, primary)
relevance slope > 0	13 / 16, one-sided Wilcoxon $p = 0.019$	11 / 16, one-sided Wilcoxon $p = 0.037$
compression slope < 0	10 / 16, one-sided Wilcoxon $p = 0.149$	10 / 16, one-sided Wilcoxon $p = 0.047$

The $\lambda = 0$ regime gave a slightly stronger relevance signal by mouse count (13 vs. 11); $\lambda = 1$ slightly sharpened compression. The dedicated input-axis-specific $\lambda = 0$ vs $\lambda = 1$ contrast is reported in the next subsection.

Target-conditioned relevance. Egocentric distances to the objects largely determine position in the arena, so part of $I(Z; Y_{\text{alloc}})$ could be shared with X_{ego}. We therefore estimated $I(Z; Y_{\text{alloc}} \mid X_{\text{ego}}) = I(Z, X_{\text{ego}}; Y_{\text{alloc}}) - I(X_{\text{ego}}; Y_{\text{alloc}})$ with the relevance-axis decoder (logistic regression onto the 10×10 bin, confusion-matrix MI), bias-corrected by replacing Z with its five shuffles while keeping X_{ego} intact. Z carried position information beyond the egocentric state in 59 of 61 sessions (mean 0.24 bits), and this conditional part grew from 0.20 bits on day 1 to 0.30 bits on day 4, rising in 11 of 16 mice (one-sided Wilcoxon $p = 0.072$).

S9.4 Input-axis $\lambda = 0$ vs $\lambda = 1$ contrast

The hippocampus input-axis quantity $I(X_{\text{ego}}; Z)$ is computed on the AE latent Z, so its trajectory in principle depends on how the AE allocates the fixed-capacity bottleneck between reconstruction and

target prediction. At $\lambda = 0$ the supervised head is removed from the loss, eliminating Y-pressure on Z; comparing the input-axis trajectory between $\lambda = 0$ and $\lambda = 1$ on the full hippocampus cohort tests whether supervised allocation drives the decline.

Cohort-mean trajectory of bias-corrected $I(X_{\text{ego}}; Z)$ per day (drop-3 cohort: $n = 16$ on days 1–2; $n = 15$ on day 3 after dropping H33; $n = 14$ on day 4 after dropping H32 and H06):

λ	day 1 ($n = 16$)	day 2 ($n = 16$)	day 3 ($n = 15$)	day 4 ($n = 14$)
0 (recon-only)	1.057 ± 0.118	1.041 ± 0.157	0.988 ± 0.132	0.860 ± 0.175
1 (Y-supervised)	1.235 ± 0.126	1.117 ± 0.142	1.145 ± 0.167	1.009 ± 0.167

Per-mouse paired delta (day 4 − day 1) on the drop-3 paired cohort ($n = 14$ mice with both day-1 and day-4 sessions retained; H32 and H06 are excluded because their day-4 sessions are dropped):

λ	metric	n	mean	n_{negative}	Wilcoxon p (two-sided)
0	$\Delta I(X_{\text{ego}}; Z)$	14	-0.223	8/14	0.326
1	$\Delta I(X_{\text{ego}}; Z)$	14	-0.271	11/14	0.068

At $\lambda = 0$ the AE has no Y in the loss and no mechanism for Y-driven capacity reallocation, yet the input axis $I(X_{\text{ego}}; Z)$ fell with the same direction and similar magnitude as at $\lambda = 1$ (mean drop -0.22 vs. -0.27 bits; 8/14 vs. 11/14 mice declining). The two trajectories were essentially parallel. The paired d1→d4 contrast at $n = 14$ paired drop-3 was underpowered (Wilcoxon two-sided $p = 0.068$ for $\lambda = 1$); the slope test on all four days at $n = 16$ (Section S11, drop-3, Wilcoxon $p = 0.047$) and the per-component analysis (d_{obj2} at $p = 0.019$, all four distances declining on average) provided convergent evidence at higher resolution. The input-axis compression therefore reflected structure in T itself, faithfully captured by the recon-only AE; the recon-only Z has no incentive to discard input information, and the fact that it did so means the input information was becoming harder to extract from T.

A second, convergent line of evidence came from per-component selectivity: compression localized to the egocentric distances $d_{\text{obj1..4}}$, while speed, angular velocity, and head-direction components were flat throughout. A uniform-capacity-pressure mechanism cannot be component-selective; only the d_{obj} components declined because only they were encoded in Z at day 1 to begin with.

Scope. This contrast tests the role of Y-supervision specifically. It leaves open the more general question of whether a fixed-dimensional bottleneck on its own could shift the fraction of X captured in Z across days, e.g. if the effective dimensionality of T varies day-to-day for reasons unrelated to X. The per-component selectivity reported above narrows this: a uniform capacity-pressure mechanism would affect all components equally, whereas only the egocentric-distance components decline.

S9.5 Per-mouse direction summary

To check that the cohort-mean motion reflected coherent per-animal behavior rather than averaging over heterogeneous trajectories, we classified each mouse's day-1→day-4 motion into a 3×3 grid by sign of the per-axis slope (decrease / near-zero / increase).[2] Counts per cell, both regimes:

	$\lambda = 1$			$\lambda = 0$		
Y direction	X decrease	X near	X increase	X decrease	X near	X increase
Y increase	3	5	3	5	3	4
Y near	2	0	0	0	0	1
Y decrease	3	0	0	3	0	0

The relevance axis rose in 11/16 mice at $\lambda = 1$ and 12/16 at $\lambda = 0$, the plurality direction in both regimes. The X-increase × Y-decrease cell was empty (0/16) in both regimes; a uniform null over

[2]The per-axis slope-sign classification used here partitions per-mouse outcomes into three buckets to expose joint (X, Y) structure; the cohort slope test in Section S9 (Wilcoxon on continuous slope) gives 13/16 mice with relevance-axis slope > 0 at $\lambda = 0$, vs. 12/16 here, the difference being one mouse with positive but small slope binning into the "near-zero" bucket.

the nine cells at $n = 16$ would expect ≈ 1.8 mice per cell, so an empty diagonally opposite quadrant is consistent with a coherent direction across animals rather than nine-way scatter.

S9.6 Better and worse learners

Does an animal's position in the information plane relate to how well it learned? Free exploration has no task score, so we related the two coordinates across the 14 mice with a retained day-4 session. Day-4 retention $I(X_{\text{ego}}; Z)$ and day-4 relevance $I(Z; Y_{\text{alloc}})$ were associated across these mice (Spearman 0.70, $p = 0.005$; 0.75 with the number of matched neurons partialled out), and day-4 retention was associated with day-4 arena-coverage entropy, a quantity taken from tracking alone (Spearman 0.71, $p = 0.004$). A median split on day-4 relevance (7 against 7) gives the same picture (Supplementary Fig. S6); because the ranking variable is the relevance coordinate itself, the groups are compared on retention only. On day 1 the groups did not differ on either axis (two-sided Mann–Whitney, retention $p = 0.46$, relevance $p = 0.38$). By day 4 the better learners kept more egocentric information (median 1.31 against 0.53 bits, $p = 0.007$; $p = 0.028$ after Holm correction across the four days, with day 4 fixed by the design). The longitudinal contrast went the same way: from day 1 to day 4 retention changed by a median of 0.00 bits in better learners and -0.30 bits in worse ones ($p = 0.007$; four-day slope, $p = 0.018$). We read this analysis as outcome-linked and diagnostic, not as a temporal forecast.

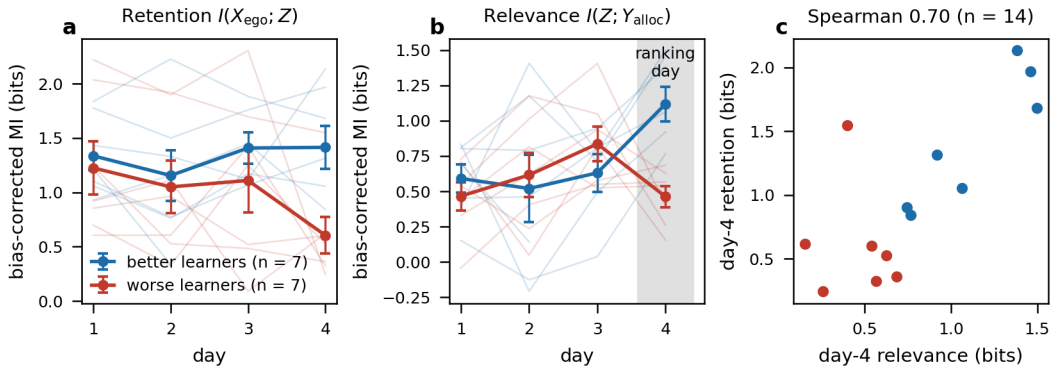

Figure S6: Better and worse learners in the hippocampus. **a.** Retention $I(X_{\text{ego}}; Z)$ per day for mice ranked above (blue) and below (red) the median of day-4 relevance; thin lines are single mice, thick lines group means with SEM. **b.** Relevance $I(Z; Y_{\text{alloc}})$ for the same groups; day 4 (shaded) is the ranking variable and is not a test. **c.** Day-4 retention against day-4 relevance, one point per mouse, coloured by group.

S10 Hippocampus bias correction mechanism

As mice familiarized with the arena, spatial occupancy concentrated and the relevance-axis target Y_{alloc} lost entropy. This drove the naive scalar metrics — above-chance decoder accuracy and raw I_{raw} — opposite to the encoding signal, because both depend on the $H(Y_{\text{alloc}})$ ceiling that deflated with familiarization. Bias correction by shuffle subtraction recovers the underlying refinement. We document the mechanism, then summarize per-mouse direction across the cohort.

S10.1 Coverage shrinkage

$H(Y_{\text{bin}})$ cohort means shrank with familiarization:

day	$H(Y_{\text{bin}})$ bits	$n_{\text{bins,visited}}$ (max 100)	mean distance
1	5.34	∼74	73.5
2	4.82	∼64	50.2
3	4.27	∼63	49.3
4	4.29	∼58	48.2

Total cohort-mean drop -1.05 bits ($\approx 20\%$) on the spatial coverage entropy of the test frames; cohort distance fell 34%. Both I_{real} and I_{shuf} are computed from the same finite sample with the same Z and Y marginals; only the temporal binding between them differs. The finite-sample bias of the MI estimator depends on the marginals and sample size, both shared between the two estimates, so the bias cancels in the difference $I_{\text{corr}} = I_{\text{real}} - I_{\text{shuf}}$. Coverage shrinkage drives both raw and shuffle estimates downward through the $H(Y_{\text{alloc}})$ marginal but does not affect the true joint dependence; the corrected difference therefore tracks the encoding signal rather than coverage drift.

S10.2 Why raw decoder accuracy gives the wrong sign

Occupancy concentration inflated both raw decoder accuracy (majority-class prediction becomes trivially stronger) and the majority-class chance baseline. The chance baseline rose faster, so the above-chance gap was flat-to-negative cohort-wise:

day	raw acc (LogReg)	majority-class chance	acc − chance
1	0.103	0.105	-0.002
2	0.092	0.147	-0.054
3	0.133	0.246	-0.113
4	0.155	0.219	-0.064

Raw accuracy rose cohort-wise ($0.10 \rightarrow 0.15$) and naively reads as the decoder improving with familiarization; chance itself rose faster ($0.10 \rightarrow 0.22$), and the above-chance gap was flat-to-negative. As a per-mouse cross-check, the slope of above-chance accuracy on the AE latent tracked the slope of corrected $I(Z; Y_{\text{alloc}})$ across animals (Spearman $r = +0.49$, $p = 0.055$, $n = 16$ mice; MLP variant), while naive raw accuracy did not ($r = -0.14$). The bias-corrected MI (real − shuffle) recovered the encoding signal that the scalar accuracy metrics hide.

S10.3 Cohort-mean MI trajectory at $\lambda = 1$

Drop-3 cohort means at day 1 and day 4 (61 sessions, 16 mice; deltas are differences of the cohort means):

quantity	day 1	day 4	Δ
$H(Y_{\text{alloc}})$	5.34	4.29	-1.05
I_{raw}	2.34	1.79	-0.56
I_{shuffle}	1.81	1.00	-0.81
$I_{\text{corrected}}$	0.54	0.79	$\mathbf{+0.25}$

Both I_{raw} and I_{shuffle} deflated; I_{shuffle} deflated faster because the shuffle signal lives entirely on autocorrelation-induced bias whose fuel (the bin-overlap structure on which the bias depends) shrinks with fewer occupied bins. Subtraction isolates the stable encoding component. As a sanity check, $I_{\text{raw}}/H(Y_{\text{alloc}})$ stayed at $0.44 \rightarrow 0.42$ (raw signal tracked the ceiling) while $I_{\text{shuffle}}/H(Y_{\text{alloc}})$ dropped $0.34 \rightarrow 0.23$ (bias lost fuel).

S11 Hippocampus sensitivity analyses and confounder controls

S11.1 Aggregate session-inclusion sensitivity

The aggregate compression direction is the per-mouse linear slope of $I(X_{\text{ego}}; Z)$ across the four days, summarized across the cohort by one-sided Wilcoxon signed-rank against zero. Its p-value is sensitive to session-inclusion choice: we varied the choice across four session sets.

The cohort definition is drop-3 (Section S1): the three sessions with severe freezing (rest or freezing in 87–96% of the session; distance traveled 11.8, 13.9, 17.3 arena lengths) were excluded; the next-lowest sessions by distance, H06 day 3 at 18.4 and H39 day 4 at 18.6, were kept. Beyond these, distance jumped to 31 for the rest of the cohort, so the lowest five sessions formed a contiguous low-activity tail.

We compared drop-3 with three alternative cuts: "no exclusion" (use all 64 sessions), "drop-4" (also H39 day 4, the fourth session with effective X_{ego} dimensionality ≤ 3; Section S1) and "drop-5" (the cleanest distance-gap cut, excluding the entire low-activity tail below the gap of ~ 12 arena lengths to the rest of the cohort). The aggregate test is per-mouse linear slope of $I(X_{\text{ego}}; Z)$ summarized by one-sided Wilcoxon signed-rank:

exclusion	criterion	n_{excl}	slope$_X < 0$	Wilcoxon p
no exclusion	all 64 sessions	0	10/16	0.037
drop-3 (cohort def)	severe freezing	3	10/16	0.047
drop-4	also H39 day 4	4	9/16	0.106
drop-5	all sessions below that gap	5	8/16	0.334

Drop-3 kept all 16 mice, each with at least three days; the result $p = 0.047$ was essentially identical to the no-exclusion result ($p = 0.037$) since the mean slope changed little (-0.103 to -0.089), although single affected mice changed more (H32 from -0.37 to -0.20). Drop-5 over-excluded: four mice (H32, H06, H33, H39) lost 1–2 days each, reducing their slopes to 2–3 day fits, and the cohort signal collapsed to $p = 0.334$. The principled choice is the freezing-based drop-3, which leaves every mouse a slope over at least three days; choosing the larger distance gap as the cut would conflate low-activity sessions with freezing artifacts and discard mouse-level evidence. The relevance-axis rise depends on the same choice in the opposite way: its per-mouse slope was positive in 10/16 mice with no exclusion (one-sided Wilcoxon $p = 0.096$), 11/16 with drop-3 ($p = 0.037$), 12/16 with drop-4 ($p = 0.011$) and 12/16 with drop-5 ($p = 0.001$); at $\lambda = 0$, in 11/16 ($p = 0.072$), 13/16 ($p = 0.019$), 14/16 ($p = 0.012$) and 14/16 ($p = 0.001$).

S11.2 Per-component slopes

The aggregate $I(X_{\text{ego}}; Z)$ averages over six components with very different day-1 baselines. Per-component cohort slope ($\lambda = 1$, $n = 16$, drop-3); we additionally included head direction and time-since-start as sanity-check variables outside X_{ego}:

variable	day-1 $I(Z; v)$	mean slope (bits/day)	n declining	one-sided p	verdict
speed	0.015	-0.000	6/16	0.609	flat
ω	0.001	$+0.001$	8/16	0.783	flat
head direction	0.004	$+0.011$	6/16	0.863	flat
d_{obj1}	0.355	-0.027	11/16	0.065	decline
d_{obj2}	0.383	-0.040	13/16	**0.019**	decline
d_{obj3}	0.315	-0.009	9/16	0.264	trend
d_{obj4}	0.388	-0.030	10/16	0.137	trend
time-since-start	-0.097	-0.021	10/16	0.248	n.s.

Of the components of X_{ego}, only the egocentric distances $d_{\text{obj1..4}}$ declined on average (d_{obj2} reaching significance, the other three with the same sign). Speed, angular velocity and head direction were flat throughout; their day-1 cohort $I(Z; v)$ was near zero, so they were not encoded in Z at day 1 and there was nothing to compress. Time-since-start had a negative corrected value at day 1 and a non-significant slope. Head direction and time-since-start are not part of X_{ego} but were tested here to rule out a uniform-encoding mechanism: if Z were a junk-bag that compressed all variables uniformly, these two non-task variables would also decline. None of the eight variables showed a significant rise; the input-axis decline was therefore selective for the encoded egocentric-distance subspace. These slopes include H39 day 4 and H06 day 3, where three and one object distances are constant. With those objects placed from the mouse's other sessions and Z unchanged, d_{obj2} still declined in 13 of 16 mice ($p = 0.019$), aggregate retention in 10 of 16 ($p = 0.052$), and day-4 retention and relevance were associated at Spearman 0.64 across the 14 mice ($p = 0.013$). Without these two sessions d_{obj2} declined in 11 of 15 mice ($p = 0.126$).

The pattern was preserved under unsupervised training ($\lambda = 0$): all four distance slopes were negative (-0.010 to -0.031 bits/day; smallest one-sided $p = 0.137$, d_{obj3}), and speed, ω, head direction, and time-since-start remained flat. Magnitude was weaker than at $\lambda = 1$ but the per-component selectivity pattern was the same.

S11.3 Raw-calcium decoder confirms speed/ω/head direction are not in T

The flat per-component slopes for speed, angular velocity, and head direction could in principle reflect an artifact of the probe (these signals are present in T but the probe does not encode them). To rule this out, we tested whether the signals were predictable from *raw* calcium activity at all. For each of the 64 sessions, we fit a Ridge regression of each behavioral target on the raw calcium population vector of all recorded neurons at the full frame rate (no AE), with a contiguous 70/30 train/test split as in the main pipeline. Targets: speed, angular velocity, head direction, and d_{obj1} as a positive control.

target	median R^2 (test)	n_{sessions}
speed	-1.97	64
angular velocity	-1.73	64
head direction	-1.87	64
d_{obj1}	$+0.16$	64

Speed, angular velocity, and head direction were not predictable from raw CA1 activity (median $R^2 \ll 0$, i.e. predictions worse than the session mean). This pattern is consistent with predominant encoding upstream of CA1 — speed in medial entorhinal cortex [Kropff et al., 2015], head direction in postsubiculum [Taube et al., 1990] and thalamus [Taube, 1995]. The object-distance positive control was modestly predictable (median $R^2 = +0.16$). The flatness of speed, angular velocity, and head direction in the per-component analysis above therefore reflected the absence of these signals from T rather than a property of the probe.

S11.4 Per-session GLM: what does Z encode?

To check what the autoencoder latent encodes, we computed per session the squared correlation r^2 between each of the 16 latent dimensions and each of 8 behavioral variables — allocentric position (x, y), the four egocentric object distances $d_{\text{obj1..4}}$, speed, and angular velocity ω — on the test frames, and subtracted the same quantity computed on the latents of the five autoencoders retrained on rolled activity (Section S3.1); latent dimensions were ranked per variable by r^2, separately for real and shuffled latents, before averaging. Head direction was excluded from the GLM because Section S11.3 establishes it is not predictable from raw CA1 activity (median $R^2 = -1.87$ across sessions); including a predictor unrepresented in T would inflate degrees of freedom without adding identifiable signal. Cohort means below are taken across the 61 sessions and the 16 latent dimensions, at $\lambda = 1$ (primary) and $\lambda = 0$:

predictor	mean Δr^2, $\lambda = 1$	max, $\lambda = 1$	mean, $\lambda = 0$	max, $\lambda = 0$
x	0.063	0.198	0.047	0.175
y	0.068	0.197	0.056	0.167
d_{obj1}	0.049	0.147	0.040	0.123
d_{obj2}	0.049	0.152	0.039	0.112
d_{obj3}	0.052	0.140	0.044	0.106
d_{obj4}	0.059	0.147	0.046	0.116
speed	0.002	0.007	0.002	0.007
ω	0.000	0.001	0.000	0.001

Allocentric position (x, y) was the primary encoded signal; the four egocentric object distances were secondary; speed and ω were indistinguishable from the shuffle null (mean $\Delta r^2 < 0.01$; at most 2 of 16 ranked latent dimensions reached $p < 0.001$, versus 11–16 of 16 for x, y, and $d_{\text{obj1..4}}$); the full per-(latent, predictor) heatmap appears in Figure S7. This was consistent with the per-component compression pattern (Section S11): the components that the GLM identified as encoded in Z were exactly the ones that declined across days, while speed and ω — which were not in Z to begin with, and not even predictable from raw T above — had nothing to compress. The same ordering held under unsupervised training ($\lambda = 0$, right columns of the table), so allocentric encoding was not a consequence of supervised training against the relevance target.

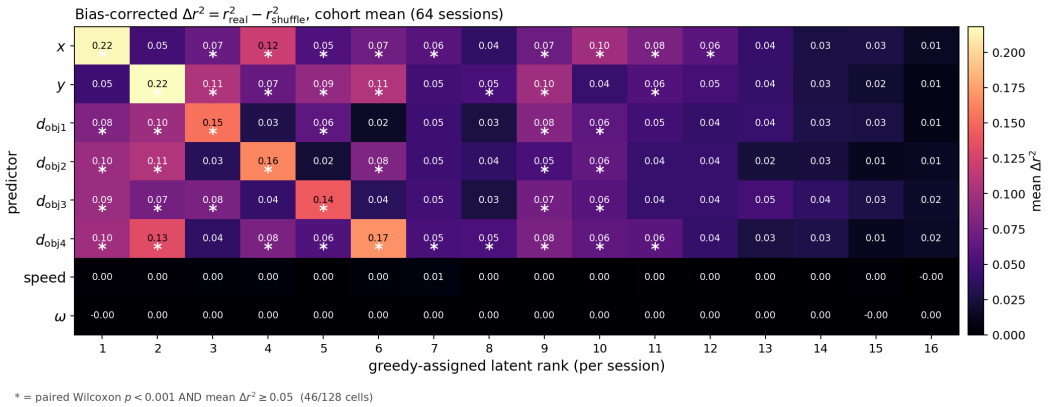

Figure S7: Per-session GLM cohort summary: bias-corrected $\Delta r^2 = r^2_{\text{real}} - r^2_{\text{shuffle}}$ for each (latent dimension, predictor) pair, averaged across the 64 sessions of an earlier run of the pipeline on all recorded neurons at the full frame rate and $\lambda = 0$. Latent dimensions are greedy-assigned to predictors in rank order. Asterisks mark cells with paired Wilcoxon $p < 0.001$ and mean $\Delta r^2 \geq 0.05$ across sessions.

S11.5 Relevance-axis day-coefficient under behavioral-habituation controls

The relevance axis $I(Z; Y_{\text{alloc}})$ rose with experience, but mice also habituated behaviorally over the same days (61-session cohort, day 1 to day 4): speed dropped by 19%, distance traveled by 34%, the coverage entropy $H(Y_{\text{alloc}})$ of the whole session by 1.0 bit ($5.99 \rightarrow 4.98$), and the fraction of session at rest or freezing rose from 14% to 36%. We tested whether the day-by-day rise was independent of habituation by fitting a linear mixed-effects model on the 61 sessions with the day index plus seven behavioral covariates as fixed effects and a per-mouse random intercept. Let $\mathbf{B}$ collect the six locomotion-related covariates — mean speed, locomotion fraction, rest/freeze fraction, rear fraction, object fraction, and distance traveled. The model is:

$$I(Z; Y_{\text{alloc}}) \sim \text{day} + H(Y_{\text{alloc}}) + \mathbf{B} + (1 \mid \text{mouse}).$$

Day-only model: day coefficient $= +0.086$ bits/day, $p = 0.034$. Day and $H(Y_{\text{alloc}})$ only: $+0.139$ bits/day, $p = 0.004$. Full model:

term	estimate (bits/day)	SE	p
intercept	-1.857	1.036	0.073
day	$\mathbf{+0.105}$	$\mathbf{0.054}$	$\mathbf{0.052}$
$H(Y_{\text{alloc}})$	$+0.324$	0.150	0.030
mean speed	-0.937	1.286	0.466
locomotion fraction	$+0.395$	1.145	0.730
rest/freeze fraction	$+1.545$	0.927	0.096
rear fraction	$+0.238$	1.391	0.864
object fraction	-1.623	0.811	0.045
distance traveled	$+0.005$	0.006	0.413

Holding $H(Y_{\text{alloc}})$ constant raised the day coefficient from $+0.086$ to $+0.139$ bits/day; with all seven covariates it was $+0.105$ bits/day, still above the day-only value. The $H(Y_{\text{alloc}})$ coefficient was positive ($+0.324$, $p = 0.030$): occupancy concentration suppressed the apparent day effect in the day-only model. Of the locomotion covariates, object fraction reached $p = 0.045$ and rest/freeze fraction $p = 0.096$; the others had $p > 0.4$. Behavioral habituation was therefore masking the encoding refinement, not driving it.

The drop-3 exclusion (Section S1) and the rest/freeze-fraction covariate above operate at non-overlapping scales: drop-3 removes the extreme-freezing tail (rest or freezing in 87–96% of the session) at the session-inclusion level, while the covariate models the remaining continuum across the retained sessions. They do not double-count.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the shared analytical scaffold applied to three substrate-different systems and the directional findings, with scope caveats discussed in the Discussion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are addressed in the Discussion (cohort size, paradigm heterogeneity, and the absence of dense temporal sampling in the biological systems).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical and contains no theorems requiring proofs; estimator assumptions are stated where used (Section 2.3).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The shared scaffold, per-system datasets, estimators, cross-validation, and bias correction are specified in Section 2; full hyperparameters and shuffle-null protocols are in Supplementary §S2–§S3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: CIFAR-10 and the Goltstein cortex dataset are public (cited). Analysis code is public under the MIT license (`https://github.com/niveousdragon/ib-neurips2026-code`), with cached intermediate results that reproduce the figures and statistics of the paper without the raw data. The hippocampal calcium recordings are at DANDI (`https://dandiarchive.org/dandiset/002047`). The SNN activity across the 26 epoch checkpoints, summed over timesteps, the MEI file and the learning curves are at Zenodo (`https://doi.org/10.5281/zenodo.23962331`, CC BY 4.0). The SNN training script is `snn/SpikingJelly/sj_snn_train.py` in `https://github.com/iabs-neuro/mango`, commit `8b96b583616a660c11f9734cf15d20317ed252c7`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Per-system data splits, optimizer, bottleneck dimension, and CV scheme are in Sections 2.2 and 2.3; full hyperparameter table in Supplementary §S2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All cohort-level claims report Wilcoxon signed-rank p-values with n stated; figures show SEM error bars per cohort or per-mouse 95% CIs as labeled in captions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All analyses run on a single workstation with one consumer GPU; per-session AE training and probes take seconds to a few minutes. Total compute and per-stage runtimes are summarized in Supplementary §S2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public ML benchmarks (CIFAR-10), the publicly-available Goltstein et al. (2021) cortex dataset, and newly-collected hippocampal calcium recordings performed under the host institution's approved animal-research protocols (Supplementary §S1); no human subjects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational analysis bridging representation learning and systems neuroscience with no direct application path to deployment that would warrant a societal-impact discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No generative models or scraped data are released; the analysis code and small mouse-imaging dataset pose no foreseeable misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All assets are cited at first use and consumed in accordance with their license terms:

- CIFAR-10 [Krizhevsky, 2009]: no explicit formal license; publicly available for research use as documented at the dataset's distribution page.
- Goltstein et al. (2021) cortex dataset [Goltstein et al., 2021]: CC-BY-SA 4.0; deposited at G-Node GIN, repository `pgoltstein/category-learning-visual-areas`. The ShareAlike clause applies to redistribution or adapted versions of the dataset itself, not to analytical work that cites it; standard academic citation suffices for our use.
- SpikingJelly [Fang et al., 2023]: Open-Intelligence Open Source License V1.0.
- CaImAn [Pnevmatikakis et al., 2016]: GPL v2 or later.
- NoRMCorre [Pnevmatikakis and Giovannucci, 2017]: GPL v2 or later.
- CellReg [Sheintuch et al., 2017]: CC BY-NC-ND 4.0.

The assets introduced by this work are public: the hippocampal calcium dataset (DANDI, `https://dandiarchive.org/dandiset/002047`), the SNN activity (SpikingJelly ResNet-18 forward passes on CIFAR-10 across the 26 training-epoch checkpoints used in this paper, summed over timesteps; Zenodo, `https://doi.org/10.5281/zenodo.23962331`, CC BY 4.0), and the analysis code with its cached intermediate tables (`https://github.com/niveousdragon/ib-neurips2026-code`, MIT license). Autoencoder weights were not stored; the released scripts retrain the autoencoders, and the cached results reproduce the figures.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The analysis code is public with author attribution under the MIT license (`https://github.com/niveousdragon/ib-neurips2026-code`), with cached intermediate tables that reproduce the figures and statistics of the paper. The new assets are the hippocampal calcium dataset (DANDI, `https://dandiarchive.org/dandiset/002047`) and the SNN activity (Zenodo, `https://doi.org/10.5281/zenodo.23962331`, CC BY 4.0). Autoencoder weights were not stored; the released scripts retrain the autoencoders. Documentation accompanies each asset: experimental and preprocessing protocol for the hippocampal recordings (Section 2 and Supplementary §S1); training, hyperparameter, and bias-correction details for the SNN and autoencoder artifacts (Sections 2.2 and 2.3 and Supplementary §S2, §S3).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects; mouse experiments were performed under the host institution's animal-research approvals (Supplementary §S1).

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An LLM-based coding assistant (Claude Code, Anthropic) was used during analysis development to help write and refactor analysis code and figure-generating scripts, and to assist with manuscript editing. It did not contribute to study design, methodological choices, statistical interpretation, or scientific claims; all analysis decisions, results, and figure semantics were specified, reviewed, and validated by the authors.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.