

Reciprocal connections dynamically build consensus between neocortical areas

Received: 22 July 2025

Accepted: 5 August 2026

Published online: 18 September 2026

 Check for updatesMitra Javadzadeh ^{1,2,3,5}✉, Marine Schimel ^{1,4,5}✉, Sonja B. Hofer ³,
Yashar Ahmadian ^{1,6} & Guillaume Hennequin ^{1,6}

The neocortex is organized into specialized areas. Although computations within individual areas have been well studied, it is unclear how these regions function collectively and reconcile potential conflicts to form coherent percepts and decisions. We investigated the joint dynamics of primary (V1) and higher-order lateromedial (LM) visual areas in mice using simultaneous multi-area electrophysiological recordings along with focal optogenetic perturbations to causally manipulate neural activity. We used data-driven nonlinear system identification to construct biologically constrained latent circuit models of both areas. This approach revealed that reciprocal excitatory connections between V1 and LM implement an approximate line attractor in their joint dynamics: this selectively slows the decay of congruent activity patterns while accelerating the decay of inconsistent ones, thereby dynamically achieving inter-area consensus. This mechanism predicts different timescales for consistent versus inconsistent activity patterns across areas, which we verified in our data. These findings, together with our mechanistic theory, identify dynamic consensus building as a general principle of distributed cortical computation.

The neocortex is segregated into distinct areas that are specialized for specific functions. This organization allows for decomposing complex problems into simpler subcomputations, such as the extraction of low-level features from intricate visual scenes. However, cognition arises from the holistic integration of these processes, making it essential that the different areas work in concert and remain consistent with each other. It is unclear how such coordination is achieved and, in particular, how any conflict that might arise between local subunits can be globally resolved.

Anatomically, cortical areas are densely interconnected through reciprocal long-range inter-area connections¹, whose organization is markedly distinct from that of local circuits within a cortical area. For instance, both excitatory and inhibitory neurons have local innervation, whereas only excitatory neurons have long-range projections that may target other areas (refs. 2–4, but see refs. 5,6). The functional role of these distinct connectivity rules is not clear; it remains unknown how

excitatory inter-area connections coordinate cortical activity and unify local subunits into coherent global computations. To address this, we combined mechanistic modeling of cortical circuits with data-driven inference of circuit dynamics. This approach allowed us to build models of cortical activity that not only explained neural responses quantitatively but also captured the causal effects of optogenetic perturbations and had biologically interpretable components, including local and long-range connections, whose functional significance we could interrogate.

Results

We focused on the joint activity dynamics of the V1 and higher-order LM visual areas in mice during visual processing. We used simultaneous multi-channel recordings from V1 and LM performed while mice were presented with a 500-ms-long visual stimulus—one of two stationary gratings oriented at +45° or −45° (Fig. 1a,b). Mice were trained

¹Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Cambridge, UK. ²Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. ³Sainsbury Wellcome Centre, University College London, London, UK. ⁴Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA. ⁵These authors contributed equally: Mitra Javadzadeh, Marine Schimel. ⁶These authors jointly supervised this work: Yashar Ahmadian, Guillaume Hennequin. ✉e-mail: javadzadeh@cshl.edu; mschimel@stanford.edu

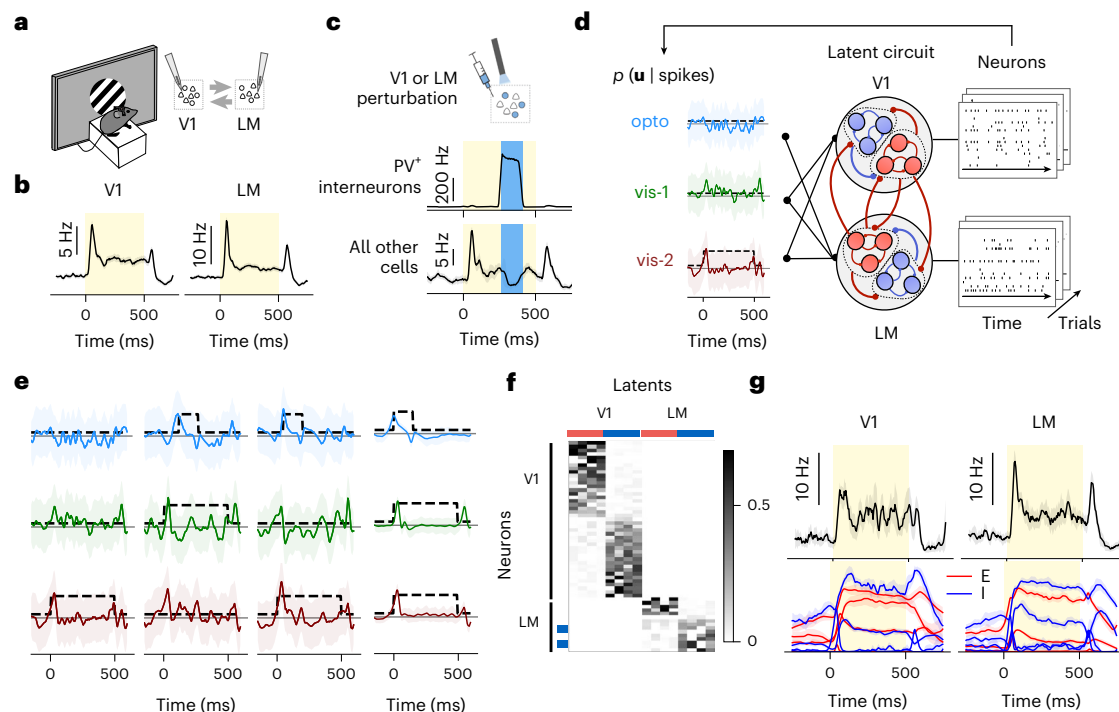


Fig. 1 | Modeling input-driven dynamics in the V1–LM network during visual processing. **a**, Head-fixed, stationary mice were presented with two differently oriented stationary grating stimuli ($+45^\circ$ or -45°), only one of which was rewarded (go/no-go task). Paired neural recordings were performed in retinotopically matched regions of V1 and LM with silicon probes in PV-Cre mice⁷. **b**, Trial-averaged and neuron-averaged spiking activity in no-go trials in V1 (left) and LM (right). A total of 194 neurons in V1 and 228 neurons in LM were recorded in seven mice across a total of 513 ± 110 (mean $\pm$ s.d.) correct trials per mouse across two stimuli. **c**, Top, in some trials, either V1 or LM was silenced through light-mediated activation of PV⁺ inhibitory cells expressing ChR2. The light onset was randomly chosen in each trial among eight different times, spanning the duration of the stimulus uniformly (at 65-ms intervals), with a total of 449 ± 98 (mean $\pm$ s.d.) silencing trials per mouse. Bottom, neuron-averaged and trial-averaged spiking activity of the optogenetically stimulated PV⁺ neurons (top) and all other neurons (bottom) in an example animal, for one silencing onset. **d**, Biologically constrained latent circuit model of V1–LM, with dynamics driven by three external inputs whose timecourse is inferred on a single-trial basis. Dashed lines indicate the time-varying standard deviations of the (zero-mean) prior distributions over these inputs. Solid lines and shaded areas indicate posterior mean and standard deviation, respectively, in one example trial, estimated from

100 posterior samples. opto, optogenetic; vis, visual; $p(\mathbf{u} | \text{spikes})$, posterior over inputs given the observed spikes. **e**, Left, inferred timecourse of inputs for three example trials. Each column shows three input channels for one trial (optogenetic perturbation in blue, go stimulus in green and no-go stimulus in red). The prior standard deviation (dashed line) tracks the presence of each input in the trial: no-go stimulus in the first trial, go stimulus paired with optogenetic perturbation in the second trial and no-go stimulus paired with optogenetic perturbation in the third trial. Shaded area is the posterior standard deviation. Right, trial-averaged timecourse of the three input channels, aligned to the input onset, shown as mean and standard deviation (shaded area) of the posterior mean across all trials where that input was present. **f**, Example readout matrix (C in equation (2)) in the fitted model, depicting the mapping from the latent units (columns, divided into two areas and into excitatory (red)/inhibitory (blue) subpopulations within each area) to the recorded neurons (rows; blue bars mark PV⁺ cells identified by optogenetic perturbations in LM). **g**, Top, average recorded activity in V1 (left) and LM (right) during the no-go visual stimulus in an example animal. Bottom, corresponding activity of the excitatory (red) and inhibitory (blue) latent units. Shaded areas around mean traces in **b**, **c** and **g** denote 95% confidence intervals (± 2 s.e.m.). All traces in **d** and **e** were smoothed with a running average of 25 ms for visualization.

to perform a go/no-go task, discriminating the two stimuli. In some trials, neural activity in either V1 or LM (varying across animals) was perturbed in brief 150-ms time windows, using optogenetic activation of inhibitory parvalbumin-expressing (PV⁺) interneurons expressing channelrhodopsin-2 (ChR2) (Fig. 1c)⁷.

We built circuit models that explicitly incorporated known aspects of cortical circuit organization, in particular the excitatory nature of long-range connections between areas and local excitation–inhibition (E–I) dynamics. In these models, the timecourse of spiking activity in V1 and LM was explained by the recurrent dynamics of the latent circuit (Fig. 1d). These dynamics were driven by time-varying inputs that we inferred for each trial, reflecting any unobserved signals external to the V1–LM circuit such as sensory or optogenetic stimuli.

Specifically, the latent circuit's activity $\mathbf{z}(t)$ evolved according to

$$\tau \dot{\mathbf{z}}(t) = -\mathbf{z}(t) + W\Phi(\mathbf{z}(t)) + B\mathbf{u}(t), \quad (1)$$

where $\tau = 20$ ms is the characteristic neuronal membrane time constant; W is the latent circuit connectivity; $\Phi(\cdot)$ is a soft-rectified nonlinear

activation function; and $\mathbf{u}$ is a set of trial-specific external input signals that enter the dynamics through the input matrix B (Methods). The activity of this latent circuit was used to describe firing rate fluctuations in the observed V1 and LM neurons, according to

$$\mathbf{r}(t) = \exp(C\mathbf{z}(t) + \mathbf{d}), \quad (2)$$

where C is a readout matrix specifying the way in which each recorded neuron relates to the latent units and $\mathbf{d}$ is a vector of constant offsets. Action potentials were modeled as Poisson processes given these time-varying firing rates.

The latent circuit was partitioned into two areas, which mapped onto V1 and LM neurons, respectively. Moreover, the recurrent connectivity matrix W was constrained so that each area was composed of separate populations of excitatory and inhibitory units, and long-range connections between the two areas originated exclusively from the excitatory units (Methods). Although we did not know the E–I identities of most of the recorded neurons, we used a specific sparsity penalty on C to discourage any nonsensical, simultaneous association

of a neuron with both E and I latent subpopulations (Methods). This soft constraint encouraged the model to learn to label each neuron as either E or I.

To fit the model, we used iLQR-VAE⁸, a method ideally suited to learning the dynamics of a circuit when the detailed timecourse of external inputs is unknown and must, therefore, be inferred in each trial. Notably, here we did have some knowledge of what input signals might have driven the circuit in a given condition and when. iLQR-VAE lets us incorporate such information in the form of condition-specific, time-varying statistical priors over the input $\mathbf{u}(t)$ in equation (1). Thus, we used three input channels reflecting the two visual stimuli and the optogenetic perturbation events. The mapping from inputs to latents, B , was constrained such that the input channel with the optogenetic perturbation prior could only target the inhibitory latents of the stimulated area for each animal. For each channel, we learned two prior variances: the higher variance was used during the time that the corresponding stimulus was on and the lower one outside those epochs (Fig. 1d, dashed lines). This encouraged the model to use larger inputs when the stimuli were present while retaining flexibility with respect to their exact timecourse. iLQR-VAE then inferred this timecourse on a single-trial basis, by computing a posterior distribution over the input signals in each channel conditioned on the observed neural data (Fig. 1d,e, solid lines).

The parameters of the model (W , B , C , $\mathbf{d}$ and the prior variances) were obtained by maximizing the likelihood of the observed spike trains (see Table 1 for hyperparameters). For each animal, we performed multiple fits starting from random initializations (Methods) and found that each fit robustly attributed a definite E or I identity to each observed neuron (Fig. 1f; see also Extended Data Fig. 1a,b). For most fits (78.24%), the model correctly labeled all of the directly photo-stimulated neurons (known to be PV⁺ inhibitory cells) as inhibitory (Fig. 1f, cyan mark); we rejected the few models where these cells were mislabeled. Finally, for each animal, we selected the model with the best goodness of fit on held-out data (Methods and Extended Data Fig. 2; see Extended Data Fig. 1 for a visualization of the parameters learned for all animals).

Fig. 2 | Circuit-constrained models capture the statistics and underlying mechanisms of neural activity. **a**, Top, trial-averaged empirical firing rates of four example neurons in one animal. Colors denote the two visual stimuli (red and green) and one silencing condition (cyan). Middle, corresponding model-predicted firing rates in individual trials (four trials per condition), given the spikes observed concurrently in all other neurons. Bottom, corresponding spike rasters. **b**, Residual log-likelihood of the predicted firing rates of active cells (test trials; Methods) for the constrained ('C', black) and unconstrained ('UC', orange) models. Each point represents one neuron ($n = 246$ neurons across $n = 7$ animals). Paired and unpaired 'C' versus 'UC' comparisons were assessed using a two-sided Wilcoxon signed-rank test and a two-sided Wilcoxon rank-sum test ($P = 6 \times 10^{-7}$ and $P = 0.23$, respectively). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. **c**, Between-animal similarity in model dynamics linearized around the stimulus period activity, calculated as an average pairwise Procrustes distance (Methods; constrained models ('C') in black and unconstrained models ('UC') in orange; two-sided Wilcoxon signed-rank test, $P = 6.7 \times 10^{-6}$). Each point represents the distance between a unique pair of animals ($n = 21$ pairs from $n = 7$ animals). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. **d**, Left, trial-averaged activity of two example neurons, obtained either from the data (control no-go trials in black and one perturbation condition in blue) or by running the learned dynamics forward given condition-specific inferred inputs and artificial perturbations (Methods; control no-go trials in gray and one perturbation condition in cyan). The top cell is directly perturbed by the optogenetic perturbation, whereas the bottom cell is only indirectly affected. Center, model predictions for the change in firing rate of neurons in response to a simulated perturbation, against the experimentally observed values (caused by optogenetic perturbations), in one example animal.

The model captures trial-by-trial variability

We first characterized how well the learned models captured single-trial activity in our recorded neurons. For each trial, we could leave one neuron out and let the model infer the timecourse of its firing rate given the activity of the other neurons (Fig. 2a). Based on this single-trial firing rate, the model then attributed a (Poisson) likelihood to each spike for that neuron. On average, this single-trial likelihood was greater than that predicted by the peristimulus time histogram (PSTH) of the same cell obtained by averaging over the other trials in the same condition (Fig. 2b, 'residual likelihood'). In other words, our latent circuits captured the spatiotemporal structure of our recordings beyond condition averages. Accordingly, our models also captured the structure of pairwise covariances in neural activity (Extended Data Fig. 3a,b). Notably, the models did not significantly suffer from the circuit constraints that we imposed; they explained the single-trial data just as well as models whose dynamics were left unconstrained (Fig. 2b, orange, and Methods).

The model infers circuit dynamics that are consistent across animals and captures key aspects of V1–LM cortical physiology

That constrained and unconstrained models explain the data equally well, despite being entirely different families of dynamical systems, raises a concern: Have our constrained models learned dynamics that really capture the mechanics of the underlying V1–LM circuit?

A first indication of faithful dynamics reconstruction is the consistency of the learned solutions across animals. We evaluated the distance between the inferred latent flow fields between pairs of animals, accounting for an arbitrary rotation of the latent state space for each animal (Methods). This analysis revealed that the dynamics learned by our constrained latent circuits were broadly consistent across animals—indeed more consistent than in unconstrained models (Fig. 2c).

As a second, stronger test of accurate dynamics reconstruction, we probed the responses of our models to internal perturbations of the inhibitory cells and compared those to the responses observed in the V1–LM data during optogenetic perturbations. Specifically, we simulated single-perturbation trials by directly providing positive input to

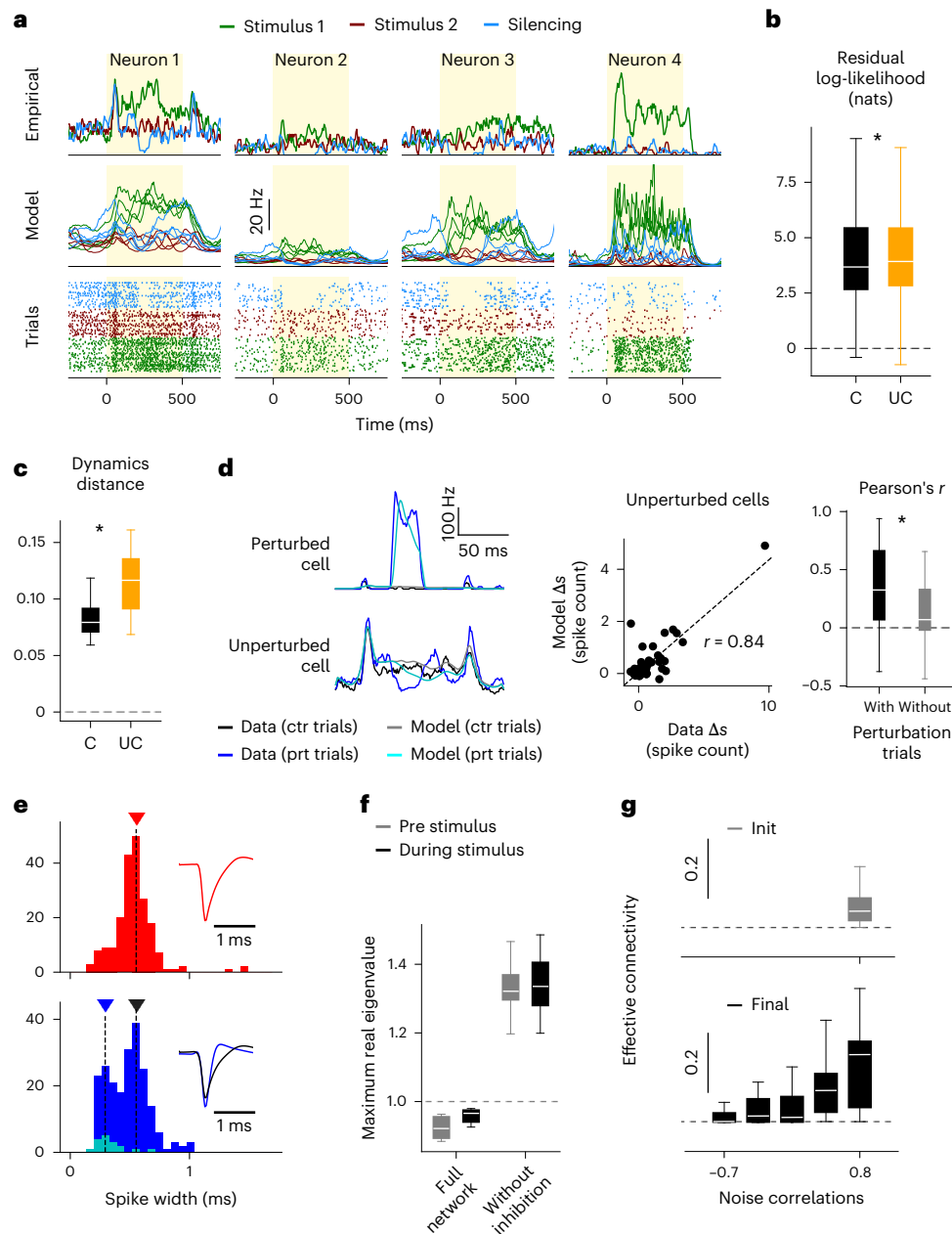
Each point represents one neuron ($n = 73$ neurons). Only neurons in V1 and LM that were not directly perturbed were included. r denotes the Pearson's correlation coefficient between the predicted and observed changes (Methods and Extended Data Fig. 3c). Right, distribution of Pearson's correlation coefficients (see middle) across animals, for models trained with perturbation data (black) and without perturbation data (gray). Each point represents one silencing condition in one animal ($n = 56$ points from $n = 7$ animals). Two-sided Wilcoxon signed-rank test, $P = 2.6 \times 10^{-4}$. Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. **e**, Histograms of spike widths for the neurons labeled by the model as excitatory (top, red) and inhibitory (bottom, blue) (known PV⁺ cells shown in cyan). Insets show the average spike waveforms (± 2 s.e.m) for those neurons lying around the marked peaks. **f**, Distribution across animals of the maximum real part of the eigenvalues of the latent circuit dynamics, linearized either before (gray) or during (black) stimulus presentation (using trial-averaged responses across both go and no-go trials). This is shown for the full model (left) and in the absence of inhibition (right) (Methods). Each point corresponds to one animal ($n = 7$). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. **g**, Effective connectivity (Methods) plotted against noise correlations in control no-go trials, for pairs of latent circuit units. This is shown at model initialization (top gray) and after training (bottom black). Each point represents a pair of latent units ($n = 112$ pairs from $n = 7$ animals). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. Init, initialization; nats, natural logarithm units; ctr, control; prt, perturbation; Δs , change in spike count between perturbation and control conditions. Asterisks in **b–d** indicate $P < 0.05$ for the statistical tests described above.

the inhibitory latent units of the relevant area for the entire duration of the photo-stimulation while replaying the external inputs inferred on a control trial (no photo-stimulation) for the relevant condition (go versus no-go). We adjusted the amplitude of each perturbation input (four parameters per animal) in order to match the trial-averaged norm of the responses of the known PV⁺ cells in the stimulated area. We then evaluated the model-predicted change in firing rates in the other neurons (per-condition averages). These predictions were positively correlated with the corresponding firing rate changes observed in the data (Fig. 2d; average Pearson's $r = 0.34$). By contrast, models trained using only no-perturbation trials failed to capture the sensitivity of the V1–LM circuit to photo-inhibition (Fig. 2d, gray; average Pearson's $r = 0.12$), highlighting the importance of optogenetic manipulations for accurate neural system identification.

As a third indication that our model has learned the correct circuit structure, we looked at the E–I identity that it assigned to each neuron in our recordings. While we used the assigned identity of the known PV⁺ cells in the photo-stimulated area as a criterion for model selection, we could study the identity assigned to the other recorded neurons.

Inhibitory neurons in the cortex are known to exhibit a bimodal distribution of spike widths: fast-spiking (PV⁺) interneurons exhibit narrow spike waveforms, whereas other (non-PV⁺) neurons have slower action potentials similar to those of excitatory neurons⁹. Consistent with this known aspect of cortical electrophysiology, we found that the neurons that the model deemed inhibitory had a bimodal distribution of spike widths (Fig. 2e). The mode of the histogram corresponding to broader spikes aligned well with the distribution of spike widths in the neurons classified as excitatory by the model.

Finally, the dynamics inferred by the model are consistent with previous studies of the mammalian visual cortex. In particular, the network operates in the inhibition-stabilized regime (Fig. 2f; refs. 10,11), whereby the excitatory subnetwork is unstable on its own but stabilized by feedback inhibition. In fact, the network is inhibition stabilized even in the absence of visual stimulation, as previously shown in mouse V1 (ref. 12). Moreover, noise correlations in the latent circuit reflect the strength of excitatory connectivity between pairs of latent units (Fig. 2g), as observed in mouse visual cortex¹³. Notably, this relationship was not present at initialization but arose after fitting the model to the data.



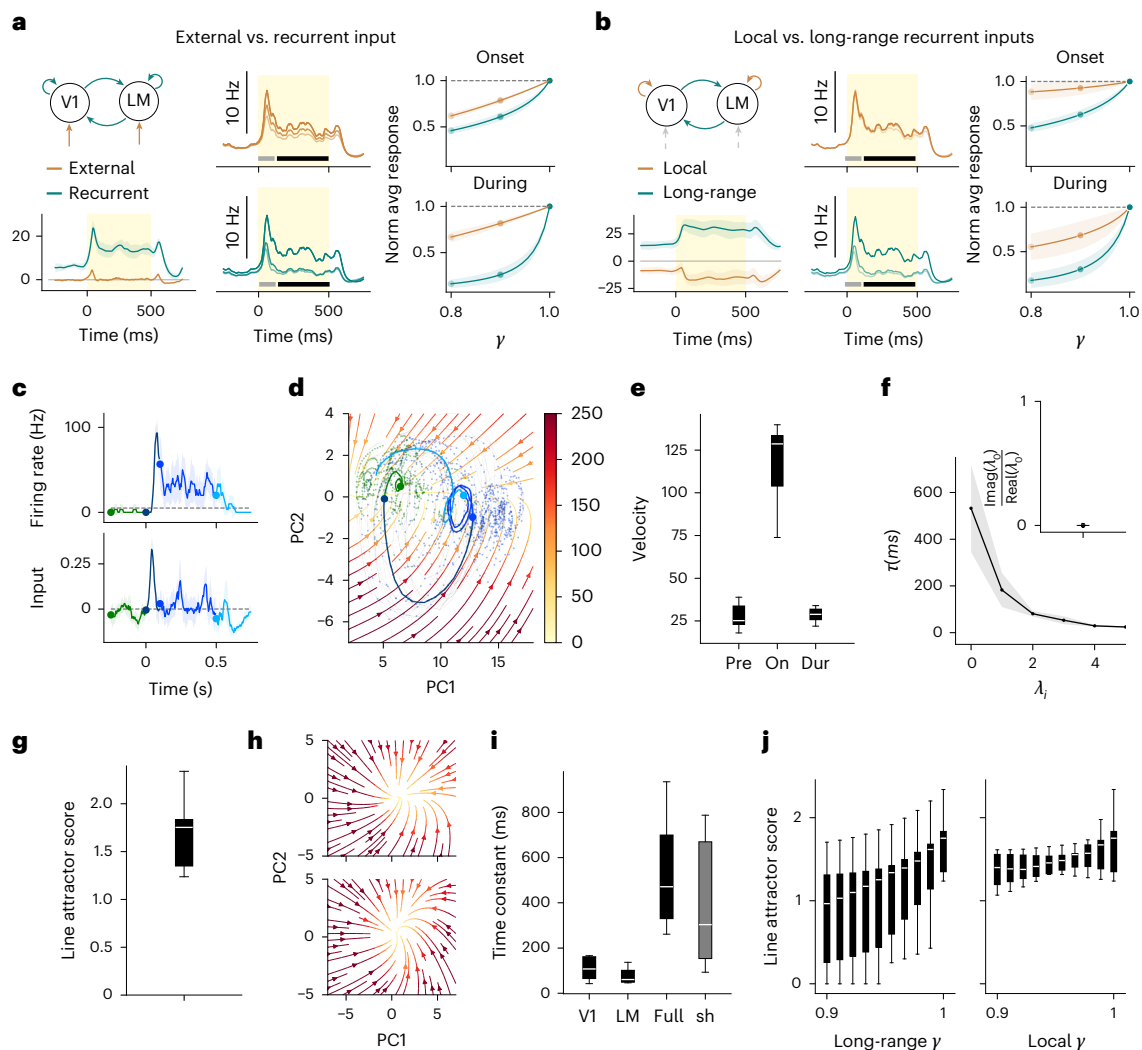


Fig. 3 | Maintenance of activity via a slow mode emerging from interacting E-I networks. **a**, Left, animal-averaged and trial-averaged external (bronze) and recurrent (teal) inputs to the V1–LM circuit. Middle top, trial-averaged and neuron-averaged network activity for an example animal, as we scale down the gain of the external inputs by 0.8, 0.9 or 1 (ordered from light to dark; Methods). The gray and black bars denote the stimulus onset and during the stimulus. Middle bottom, as in top but for recurrent inputs. Right, sensitivity—that is, change in the response of the network (Methods) as we scale down (by γ) the external (bronze) or recurrent (teal) inputs. This is shown at stimulus onset (top) and during the stimulus presentation (bottom). Shaded areas denote mean $\pm$ 95% confidence intervals. **b**, As in **a** but comparing local and long-range recurrent inputs (Methods). Shaded areas denote mean $\pm$ 95% confidence intervals. **c**, Trial-averaged firing rate of an example neuron (top) and trial-averaged inferred external input to the circuit, averaged over all latent units (bottom). The colors indicate different time segments. Shaded areas denote mean $\pm$ 95% confidence intervals. **d**, Nonlinear flow field identified for the same animal as in **c**, projected in the subspace defined by the top two PCs of the latent activity (Methods). The color bar indicates the magnitude of the velocity (before projection). The colored trajectory represents the projection of the trial-averaged inferred latent activity, with time segments color coded as in **c**. The gray trajectories (thin lines) show single-trial examples. The green points show consecutive positions along these latent trajectories at 5-ms intervals in the 250-ms pre-stimulus period (corresponding to the green section of the average trajectory). Blue points are as green points but for the 100–400-ms post-stimulus time window. **e**, Velocity of the autonomous latent dynamics (that is, latent dynamics in the absence of external input), averaged over the pre-stimulus period, around stimulus onset or during the stimulus. Each point represents one animal ($n = 7$). Box plots show the median (center line), interquartile range (box bounds) and whiskers

extending to the farthest points within 1.5 $\times$ the interquartile range. **f**, Relaxation time constants of the linearized dynamics during the stimulus (across animals; Methods). Shaded areas denote mean $\pm$ 95% confidence intervals ($n = 7$ animals). The inset shows the absolute value of the imaginary to real ratio of the eigenvalues corresponding to the slowest direction. **g**, Distribution across animals of the line attractor score (Methods) of the dynamics, linearized around the mean activity during the stimulus. Each point represents one animal ($n = 7$). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within 1.5 $\times$ the interquartile range. **h**, Flow field of the V1-only (top) or LM-only (bottom) dynamics, for the same example animal and trials as in **d**, projected in the subspace defined by the top two PCs of the latent trajectories in each area (Methods); colors are as in **d**. **i**, Distribution across animals of the slowest relaxation time constant of the dynamics in the full networks (constrained models) and the V1-only or LM-only networks ($n = 7$ animals). The gray box corresponds to the distribution of slowest time constants in networks of the size of a single area, randomly subselected from the full networks (average of 500 random selections per animal, for $n = 7$ animals). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within 1.5 $\times$ the interquartile range. **j**, Line attractor score of the V1–LM network, as we scale down the long-range (left) or within-area (right) connections by a factor γ . Each point represents one animal ($n = 7$). Box plots show the median (center line), interquartile range (box bounds) and whiskers extending to the farthest points within 1.5 $\times$ the interquartile range. Only no-go trials were used in this figure. Time windows of –400 to 0 ms, 0–100 ms and 100–500 ms were used for the pre-stimulus period, stimulus onset and during the stimulus, respectively. avg, average; Norm, normalized; On, onset; Pre, pre-stimulus; Dur, during; Imag, imaginary; PC, principal component; sh, shuffled.

Contribution of external and recurrent inputs in shaping cortical visual responses

Having established the validity of our model fits, we then used the resulting latent circuits to dissect the roles of various structural components of the V1–LM network in shaping its sensory responses. To do this, we focused on several key features of the learned latent circuit connectivity, systematically and individually downmodulated their strengths and quantified the effect of these modulations on the circuit's responses to sensory stimuli. Only no-go trials, with no reward or licking-related movement, were used for this analysis (Fig. 3).

We began by dissociating external and recurrent inputs to the latent circuit. While the average external input to each neuron was mostly transient—that is, confined to the onset and offset of the sensory stimulus—the corresponding recurrent input remained elevated for the whole stimulus duration (Fig. 3a, left), mirroring the period of sustained activity across two areas during the stimulus epoch (recall Fig. 1g). Even modest downscaling of all recurrent weights during the stimulus (Fig. 3a, center bottom, black bar) could nearly abolish these sustained responses. Similar downscaling of recurrent connectivity during stimulus onset (Fig. 3a, center bottom, gray bar) had a weaker effect (Fig. 3a, right; compare top and bottom green curves). Modulation of the external input weights had a weaker effect still (Fig. 3a, center and right), indicating that sustained activity arose primarily from recurrent connections, with external inputs triggering the onset response.

Next, we characterized the differential contributions of local versus long-range connections. Whereas the net local inputs were smaller and negative (inhibition dominated; ref. 14), the long-range inputs were stronger (and positive by design; Fig. 3b, left), leading to positive net recurrent inputs. Moreover, modulating local and long-range connection strengths separately revealed that stimulus epoch sustained activity depended strongly on long-range interactions but more weakly on within-area interactions. Together, these sensitivity analyses suggest a mechanism for sustained sensory responses in the V1–LM circuit that relies on across-area reverberation of activity, mediated by bidirectional long-range connections.

Sustained sensory responses are maintained by approximate line attractor dynamics across V1 and LM

To further characterize the origin and properties of V1–LM reverberation induced by transient inputs (Fig. 3c), we analyzed the activity flow field in the latent circuits. In the subspace defined by the two

principal components of latent activity, the autonomous flow of the latent circuit's dynamics (that is, latent trajectories obtained in the absence of external inputs) primarily converged toward a line of slow dynamics (Fig. 3d, one example mouse). Consistently across mice, the latent state trajectories underlying the neural data spent most of the stimulus epoch near this line of weak flow, only briefly leaving this region at stimulus onset and offset in response to transient external inputs (Fig. 3e). This picture is highly suggestive of line attractor dynamics^{15–18}, a regime characterized by slow decay of activity along a select direction in state space, with all other directions decaying more rapidly. Mathematical analysis of the time constants present in the latent circuits (Fig. 3f and Methods) revealed such a gap, with local dynamics around the stimulus-evoked response largely dominated by a slow mode with a timescale of approximately 400 ms, which is 20 times longer than the characteristic time constant of single neurons in our model (20 ms). Although the second-longest time constant was also slow ($\gg 20$ ms)—a point that we will return to below (Fig. 4)—it was considerably shorter than the slowest, reflected in a high 'line attractor score' (Fig. 3g and Methods).

Notably, the approximate line attractor that we identified in the latent models arose specifically from the long-range excitatory interactions between V1 and LM. First, the line attractor score was sensitive to modulation of the long-range, but not the local, connections (Fig. 3j). Second, the local dynamics in each area of the model (assessed by ablating the long-range connections) did not exhibit any line attractor (Fig. 3h) and had substantially faster dynamics (Fig. 3i). Although weaker reverberation of activity in those isolated areas could, in principle, reflect their smaller sizes, randomly thinning both areas of the full latent model (by eliminating half of their units) did yield slower dynamics than those of the isolated areas (Fig. 3i, gray). Third, single-area models fitted separately to data from each area did not exhibit a line attractor either. Not only does that corroborate the dependence of the line attractor on long-range connections established above; it also shows that our method does not automatically hallucinate a line attractor to explain slow activity fluctuations in neural recordings—a point we also verified by performing additional controls on synthetic data with known ground truth (Supplementary Information, section S2 and Extended Data Figs. 4–6; see also Discussion). Finally, consistent with the line attractor arising from reciprocal inter-area connections, recorded V1 and LM activities each exhibited faster fluctuations during inactivation of the other area (Fig. 5a).

Fig. 4 | Network time constants in minimal models of interconnected E–I circuits. **a**, In a canonical E–I circuit (top), momentary activity can always be expressed as a linear combination of two modes (middle): an 'unbalanced' mode (dashed outline) and a 'balanced' mode (solid outline). Recurrent E–I connectivity is equivalent to feedforward connectivity from the unbalanced to the balanced mode (bottom; ref. 19). The unbalanced mode exhibits fast dynamics, whereas the balanced mode can evolve more slowly depending on the degree of excitatory dominance (see **d**). **b**, Minimal E–I model of V1 and LM (top), where connectivity within each area is of the same form as in **a** and each area excites the other via long-range connections of strength ℓ . In this model, activity can be decomposed into four modes (middle): a pair of balanced and unbalanced modes in which V1 and LM activities are anti-aligned (purple, 'disagree') and a similar pair in which they align (green, 'agree'). These two pairs of modes are decoupled, and interactions within each pair are effectively feedforward (bottom). The dynamics are slowest along the mode of balanced agreement (see **e**), resulting in an approximate line attractor. **c**, This minimal model can be extended to accommodate selectivity to $\pm 45^\circ$ visual gratings, by splitting each E–I population into two differentially selective subpopulations. All connection types (local E, local I, long-range E) are composed of an unselective baseline and a selective (like-to-like) component (top). This connectivity structure gives rise to two versions (unselective, pale/selective, dark) of each of the four modes in **b** (middle) and results in slow dynamics in the two modes of balanced agreement (approximate 'plane attractor'). **d–f**, Time constants of the balanced mode(s)

for each model (colors as in **a–c**), as a function of key connectivity parameters. In **f**, only the two slow modes are shown. **g**, Flow field of the dynamics of the model in **b** in the activity plane spanned by the two balanced modes, showing convergence onto the 'agree' mode. Each line is obtained by integrating the network's dynamics starting from a different initial condition in that plane. **h**, Autocorrelation function of the neural data before (pre) stimulus (left half) and during stimulus (right half) projected onto the 'agree balanced' mode (green) and 'disagree balanced' mode (purple). These two modes correspond to the sum (respective difference) of the average V1 and LM spiking activities. Traces show the mean $\pm$ 95% confidence intervals. pre: $\Delta_{\text{max}} = 0.097$, $P = 3.2 \times 10^{-28}$, $\Delta t_{\text{slow}} = 21$ ms (Supplementary Information, section S3). during: $\Delta_{\text{max}} = 0.186$, $P = 4.2 \times 10^{-68}$, $\Delta t_{\text{slow}} = 67$ ms (Supplementary Information, section S3). P values are from a two-sided Wilcoxon signed-rank test. **i**, Same as **h** but for neural activity projected onto the pair of unselective agree/disagree modes (left) and analogous selective versions (right) defined in the main text. unselective, pre: $\Delta_{\text{max}} = 0.049$, $P = 3.3 \times 10^{-10}$, $\Delta t_{\text{slow}} = 13$ ms (Supplementary Information, section S3). unselective, during: $\Delta_{\text{max}} = 0.121$, $P = 8.2 \times 10^{-46}$, $\Delta t_{\text{slow}} = 27$ ms (Supplementary Information, section S3). selective, pre: $\Delta_{\text{max}} = 0.017$, $P = 0.0033$, $\Delta t_{\text{slow}} = 14$ ms (Supplementary Information, section S3). selective, during: $\Delta_{\text{max}} = 0.028$, $P = 7.1 \times 10^{-5}$, $\Delta t_{\text{slow}} = 26$ ms (Supplementary Information, section S3). P values are from a two-sided Wilcoxon signed-rank test. Autocorr., autocorrelation; const., constant; Dur., during; Exc., excitatory; Inh., inhibitory; Stim., stimulus.



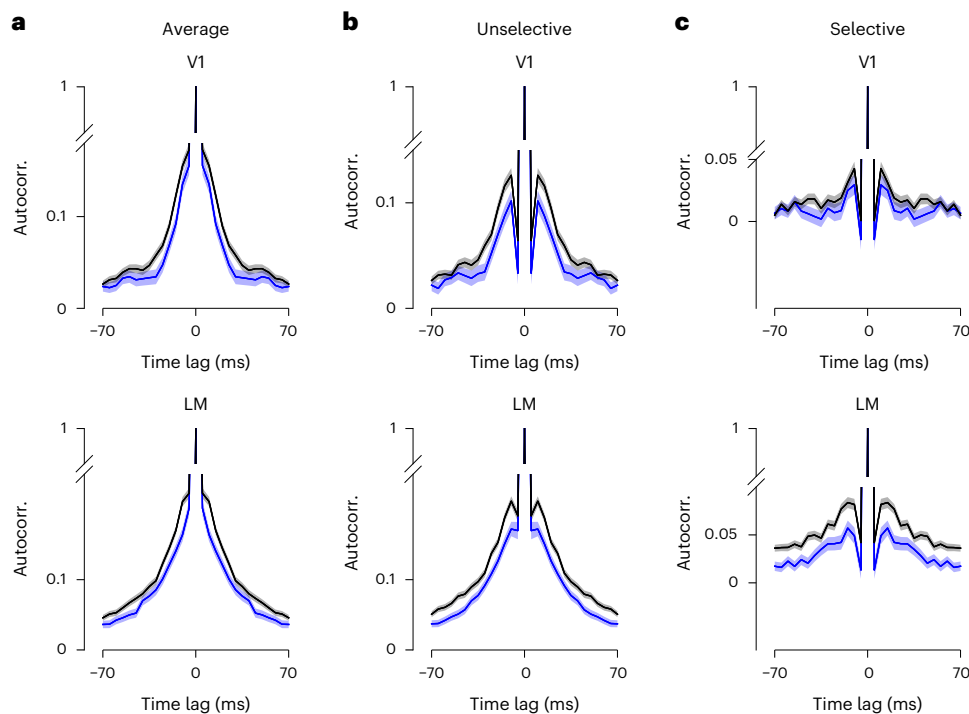


Fig. 5 | Optogenetically silencing V1 or LM decreases the contribution of slow fluctuations to the activity in the other area. **a**, Top, autocorrelation function of the average neural activity in V1 during the optogenetic silencing of LM (blue, 150-ms window) and during the corresponding window in control trials (black). Only animals with optogenetic silencing of LM were included in this analysis. Traces show mean $\pm$ 95% confidence intervals, across all silencing conditions, go and no-go trials and animals ($\Delta t_{\text{slow}} = 16$ ms; Supplementary Information, section

S3). Only correct trials were included. Bottom, as in top but for LM activity, in control trials and during V1 silencing, only including animals with V1 silencing ($\Delta t_{\text{slow}} = 25$ ms; Supplementary Information, section S3). **b**, As in **a** but for the activity projected on the unselective mode (V1: $\Delta t_{\text{slow}} = 15$ ms; LM: $\Delta t_{\text{slow}} = 17$ ms; Supplementary Information, section S3). **c**, As in **a** but for the activity projected on the selective mode (V1: $\Delta t_{\text{slow}} = 10$ ms; LM: $\Delta t_{\text{slow}} = 43$ ms; Supplementary Information, section S3). Autocorr., autocorrelation.

A minimal model of the V1–LM circuit explains the emergence of a line attractor

To understand the circuit mechanisms that underlie the emergence of a line attractor across V1 and LM, we considered simplified models of multi-area E–I networks. As a starting point, we recall a canonical model of cortical E–I circuits, with one E and one I population recurrently connected as shown in Fig. 4a (top), with E–I weight parameters e and i . In these networks, activity can be generically decomposed into two main motifs (Fig. 4a, middle): E–I imbalance (with the E population firing more than average and the I population firing less; dashed boxes) and E–I balance (both populations firing in the same way; solid boxes). In this modal decomposition, the recurrent connectivity is more easily interpreted: it acts to transiently amplify any momentary imbalance in network activity into balanced activity (Fig. 4a, ‘Schur basis’; ref. 19). Whereas E–I imbalance is typically short lived, balanced activity may linger depending on the level of excitatory dominance in the recurrent connectivity (Fig. 4a, bottom, and Supplementary Information, section S5).

Next, we extended the canonical single-area E–I model to two interacting areas, yielding an idealized reduction of our latent circuit models of V1 and LM. In this model, each area is modeled as an E–I circuit as above, and they interact via long-range excitatory connections of strength ℓ (Fig. 4b). Mathematical analysis of this model revealed a similar kind of feedforward connectivity as for single-area E–I networks, now for two different sets of unbalanced/balanced modes. In the first set, the two areas fluctuate congruently such that their patterns of E–I activities, whether balanced or unbalanced, are aligned (‘agree’, green boxes). In the other set, these patterns are anti-aligned across the two areas (‘disagree’, purple boxes). Recurrent connectivity now acts separately on each set, causing transient amplification of congruent/incongruent E–I imbalance into the corresponding balanced pattern.

Notably, long-range excitatory connectivity has an opposite effect on each set of modes: it acts to slow down activity where the two areas agree and speed up the decay of any disagreement (Fig. 4e and Supplementary Information, section S5). This separation of timescales gives rise to approximate line attractor dynamics in the combined circuit (Fig. 4g), as observed in the latent circuit models we had obtained from data (recall Fig. 3d–g). Moreover, the model clarifies that the line attractor arises specifically from long-range connections as previously shown in Fig. 3h–j. In addition, the model confirms that the line attractor score (which depends directly on the time scale separation) should grow with the strength of those long-range connections as in Fig. 3j.

Importantly, this simplified model of V1–LM interactions not only provided a qualitative explanation for the emergence of a line attractor; it also matched the dynamics of our latent circuit models quantitatively. Indeed, with optimally chosen parameters, this four-dimensional network could account for 65% of the variance in the impulse response of the (linearized) 16-dimensional network (Supplementary Fig. 1).

Notably, the simplified model of V1–LM interactions outlined above makes a prediction that can be tested independently of our latent circuit model fits. Specifically, V1–LM activity projected along the ‘balanced-agree’ mode should exhibit slower fluctuations than along the ‘balanced-disagree’ mode. To verify this prediction experimentally without relying on the latent circuit model, we estimated the contribution of each of these two modes to the momentary activity of the recorded neurons in our dataset. This was done by separately averaging the activity of V1 and LM neurons to estimate local balance in each area and then taking the sum (agree) and the difference (disagree) of these local averages. As predicted, we found that the empirical ‘balanced-agree’ mode had a longer autocorrelation decay time than its ‘disagree’ counterpart (Fig. 4h and Methods), both before (left) and during (right) the presentation of the sensory stimulus.

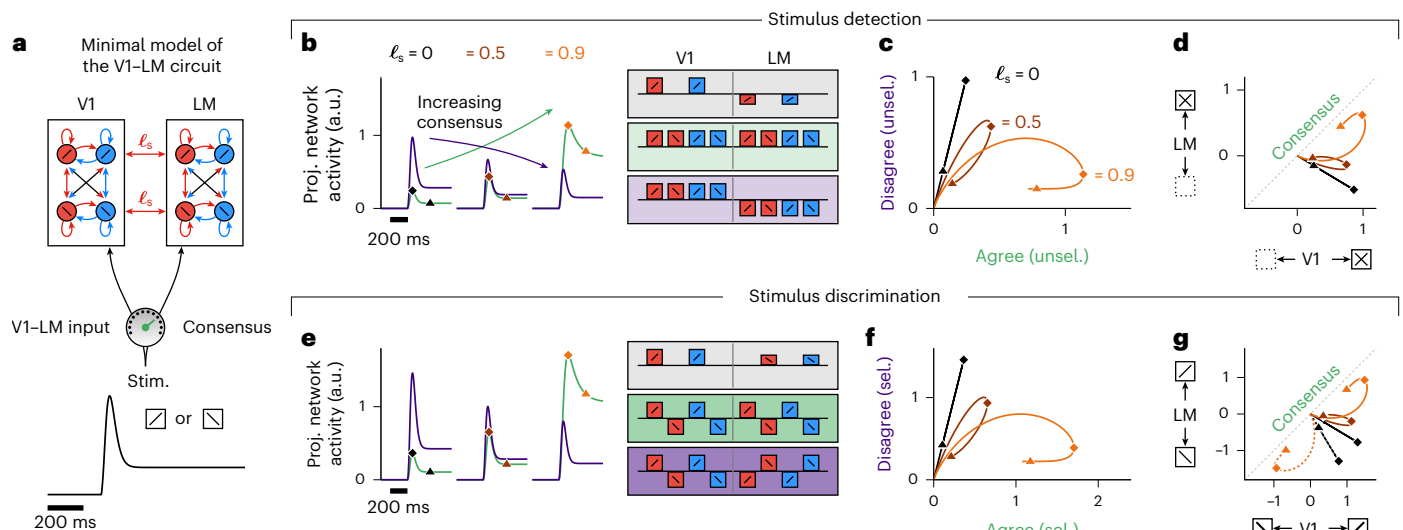


Fig. 6 | Dynamic emergence of consensus via long-range connections in minimal models of interconnected E-I networks. **a**, Schematics of the minimal selective model of the V1-LM circuit, driven by an external stimulus that can target either one of the two E-I pairs in each area depending on their orientation preference ($\odot$ versus $\ominus$), with stylized timecourse shown at the bottom displaying both transient and sustained elements. We allow for a variable degree of input coherence across V1 and LM ('input consensus' dial). **b-d**, Dynamics of V1-LM consensus for stimulus detection. **b**, Network activity projected onto the 'agree' (green) and 'disagree' (purple) modes of balanced, unselective activity (green and purple insets; recall Fig. 4b), in response to a $\odot$ stimulus that drives V1 while suppressing LM, albeit less strongly (gray inset). Response projections are shown for three values of the specific long-range connection parameter ℓ_s (0, 0.5 and 0.9), with diamond marks indicating the point of maximum consensus and triangular marks indicating 200 ms after that. **c**, Same data as in **b** with the

projection of momentary, trial-averaged network responses onto the 'agree' mode (green line in **b**) now plotted against its 'disagree' counterpart (purple line in **b**). Diamond and triangular marks as in **b**. **d**, Same data as in **b** and **c**, now showing the projections onto local unselective modes (presence versus absence of stimulus) in V1 and LM against each other. **e-g**, Same as **b-d**, for stimulus discrimination. In this case, V1 is strongly driven by a $\odot$ stimulus, whereas LM is more weakly driven by a $\ominus$ stimulus (gray inset). The relevant agree/disagree modes are now the selective modes (green and purple insets), corresponding to consensus about the identity of the stimulus rather than its presence/absence. As for detection, this conflicting stimulus gives rise to the correct consensus ($\odot$), especially for large ℓ_s . This happens even though the input itself presents more disagreement than agreement (**f**, black). Proj., projected; sel., selective; Stim., stimulus; unsel., unselective.

More generally, the model predicts slower dynamics along the balanced-agree mode compared to any other mode, including the unbalanced modes. Testing this more general prediction without referring to our latent circuit models is difficult, because estimating momentary E-I imbalance in V1 or LM directly from the neural data requires knowing the E-I identities of all cells. Nevertheless, identifying these modes based on model-predicted cell identities (see Figs. 1f and 2e) allowed us to confirm this more general prediction (Extended Data Fig. 7a).

Multi-area consensus on stimulus presence and identity via selective long-range interactions

The selective slowing down of activity patterns where V1 and LM 'agree', and concurrent quenching of patterns where they disagree, can be seen as a circuit mechanism for consensus building (Fig. 4g). We wondered about the generality of this mechanism: while the minimal two-area model of Fig. 4b gives rise to consensus regarding whether or not a stimulus is present, a similar mechanism could also underlie consensus about stimulus identity. We hypothesized that this second mode of consensus might also account for the second-slowest mode in the learned dynamics (Fig. 3f), which the simple reduced model introduced above was unable to explain.

To explore this hypothesis, we took a similar modeling approach as above. We constructed a more detailed reduced model of a two-area network (Fig. 4c) that incorporates feature specificity in its connectivity (Supplementary Information, section S5). Each area was split into two E-I subcircuits that were differentially driven by two orthogonally oriented stimuli (corresponding to go and no-go stimuli in our experiments). Recurrent E-I connectivity in each area had a degree of specificity that we could vary—that is, connectivity could be made stronger within, compared to between, the two local subcircuits

with different stimulus preference. Similarly, long-range excitatory connection strengths included both a baseline (ℓ_0) and a specific (ℓ_s) component (Methods).

The effect of the connectivity on the dynamics of this circuit could again be understood by considering a modal decomposition similar to Fig. 4b, which included (1) patterns of E-I imbalance/balance, in which (2) the two areas could either agree (green) or disagree (purple) and which (3) were either stimulus selective (dark) or unselective (light). We found that this circuit would predominantly dwell in two of these activity modes: the 'balanced-agree-selective' mode and the 'balanced-agree-unselective' mode, both of which were characterized by long time constants. As before, all 'unbalanced' and 'disagree' modes were associated with comparatively faster decay times. The slow decay of the 'balanced-agree-unselective' mode relied on strong long-range connections regardless of specificity (ℓ_0 or ℓ_s ; Fig. 4f, left), whereas the slow decay of the 'balanced-agree-selective' mode required strong specific long-range connections (ℓ_s ; Fig. 4f, right). Thus, this circuit supports the dynamic formation of a consensus across V1 and LM about both the presence of a stimulus and its identity. The presence of a second slow mode made this eight-dimensional reduced model an even better quantitative match to the linearized dynamics of the full 16-dimensional model (Supplementary Fig. 2; 77.3% of variance captured in the impulse response). Additionally, we also verified that the two slowest modes in the dynamics of our latent circuit models aligned well with the unselective and selective balanced-agree modes (Extended Data Fig. 7f).

We could again articulate the model's predictions regarding the relative timescales of these different activity modes and use our neural recordings to test these predictions. In particular, the model predicted that the network's activity should fluctuate slower along the two main modes of consensus than along the corresponding modes of

Table 1 | Model hyperparameters

Symbol	Value	Unit	Description
n	16	-	Number of latent units
m	3	-	Number of latent inputs
F_e	0.5	-	Fraction of excitatory units
τ_e	20	ms	Excitatory latents time constant
τ_i	20	ms	Inhibitory latents time constant
η	0.004	-	Learning rate
k	10	-	Scaling of the optimizer square root decay
r	0.6	-	Spectral radius of $\mathbf{W}$ at initialization
λ	1,000	-	Scale of the regularization for $\mathbf{C}$

disagreement. To test this hypothesis independently of our model fits, we estimated the degree of engagement of each neuron in the four ‘agree/disagree-selective/unselective’ modes based on its observed responses and assessed the slowness of population activity projected onto these modes. Specifically, we first extracted the sensitivity of each recorded population (V1 or LM) to the presence of a stimulus irrespective of its identity by taking the difference of its population activity vector after and before stimulus onset, denoted by $\mathbf{v}_0^{V1/LM}$ (Methods). We then defined the ‘agree-unselective’ mode as $(\mathbf{v}_0^{V1}, \alpha \mathbf{v}_0^{LM})$ and the ‘disagree-unselective’ mode as $(\mathbf{v}_0^{V1}, -\alpha \mathbf{v}_0^{LM})$, where α is a scaling factor that accounts for unequal sampling of V1 versus LM neurons in our recordings (Methods). Similarly, by computing the differences $\mathbf{v}_s^{V1/LM}$ between population responses to the go and no-go stimuli, we could define the ‘agree-selective’ and ‘disagree-selective’ modes as $(\mathbf{v}_s^{V1}, \pm \alpha \mathbf{v}_s^{LM})$, respectively.

We then projected the activity of all recorded neurons onto these four modes and computed the autocorrelations of the resulting signals (Fig. 4i). As predicted by the model, activity fluctuated slower along both ‘agree’ modes compared to the corresponding ‘disagree’ modes. This was true for activity taken both before and during the stimulus presentation. The relative slowness of the ‘agree-selective’ mode thus suggests some degree of specificity in the long-range connections between V1 and LM, consistent with experimental findings²⁰. Finally, the model predicts that slow fluctuations in both the ‘agree-unselective’ and ‘agree-selective’ modes arise specifically from closed-loop long-range interactions between V1 and LM. Consistent with this, we found that silencing one area decreased the contribution of slow fluctuations to both unselective and selective activity components in the other area (Fig. 5b,c).

Functional consequences of consensual dynamics

To explore the functional significance of slow unselective and selective consensual dynamics across V1 and LM, we revisited our minimal selective model (Fig. 6a, top) and examined its dynamics along the agree/disagree modes identified earlier. Specifically, we provided the model with a stimulus whose timecourse captured both the strong transient and weaker sustained characteristics of the input that we had inferred from the recorded spiking data (Fig. 6a, bottom). By varying the degree to which the stimulus drove (1) each area as well as (2) each subpopulation therein, we could manipulate the degree of ‘input agreement’ between V1 and LM about (1) the presence of a visual stimulus and (2) whether it is oriented at +45° or −45° (Fig. 6b,e, gray insets). We could then examine any emergent consensus in the network’s response. For example, the stimulus pattern shown in Fig. 6b (gray) drives V1 and LM in opposite directions, increasing V1 activity while suppressing LM (evidence for the presence of the stimulus in V1 and against it in LM). Mathematically, this stimulus recruits both the agree-unspecific and disagree-unspecific modes, leading to input ambiguity. In the

absence of long-range connections between V1 and LM, the network’s response directly reflects this lack of consensus in the input (Fig. 6b, left, and Fig. 6c,d, black). However, with increasingly strong long-range connections, the network selectively amplifies the input contribution to the agree-unselective mode while suppressing it for the disagree-unselective mode, thus allowing an inter-area consensus to dynamically emerge on the presence of a stimulus (Fig. 6b–d, orange). Notably, this consensus is contingent on bidirectional inter-area reverberation of activity and is substantially diminished if the feedback connections from LM to V1 are ablated (Extended Data Fig. 8). Likewise, when the input to both areas has the same total magnitude but is conflicted about stimulus orientation (Fig. 6e, gray inset), specific long-range connections contribute to the emergence of a consensus about stimulus identity (Fig. 6e–g). Notably, this consensus favors the alternative that is more strongly supported by the input (here, +45°; see Fig. 6g, dashed, for the opposite scenario). This is true even when the stimulus contributes more to the disagree-selective mode than to the agree-selective mode (as in the case shown here).

Discussion

In this study, we set out to elucidate the role of long-range connectivity in orchestrating dynamics across cortical modules. By combining data-driven and mechanistic modeling, we developed latent circuit models of observed neural activity across mouse V1 and LM, which were constrained by known properties of cortical circuit organization. These models uncovered slow reverberation of activity through long-range connections between the two areas. Further mathematical modeling revealed how this dynamical motif constrains the activity of distributed cortical modules in a way that ensures consistency of computation or ‘consensus’ between them.

Issues with model identifiability and how to mitigate them

Identifying dynamical interactions between brain areas by fitting a network model to concurrent observations of their activity is, in general, a challenging problem. Indeed, it is difficult to unequivocally tease apart external and recurrent contributions to the input that drives each neuron’s fluctuations^{8,21–23}, as neither input is directly observed. Our approach mitigates this concern in two ways. First, our dataset contained trials with optogenetic perturbations, which provide models with information about instantaneous, direct input to known cells. Second, by introducing biological constraints into the model, we not only restrict the space of possible models that fit the recorded neural activity²⁴ but also expose the model to a series of experimentally testable validation criteria, such as the correct labeling of PV⁺ cells or the effect of cell-type-specific photo-activation on network responses. Finally, our model ultimately made predictions about the relative timescales of activity in different cross-area modes, which we verified completely independently of our specific model fits (Fig. 4). Nevertheless, to mitigate some remaining concerns about identifiability, we performed additional control analyses showing that iLQR-VAE correctly rejects the line attractor hypothesis in synthetic data where slow fluctuations arise purely from slow external inputs (Supplementary Information, section S2).

Variations on the model structure do not substantially alter our results

Our data-driven models have the flexibility to learn parameters governing both the network dynamics and the external input weights, along with the timecourse of the external inputs. However, the space of possible solutions may still be constrained by key model design choices. To assess the robustness of our findings to changes in model parametrization, especially the existence of an approximate line attractor, we fitted several alternative models to the data.

First, while V1 is known to be the primary recipient of visual input (through the dorsal lateral geniculate nucleus (dLGN) pathway), our

models included external inputs to both V1 and LM. This choice was motivated by anatomical evidence: LM receives visual extracortical input from the superior colliculus via higher-order thalamus²⁵ (below, we also discuss that the inferred external inputs appear to contain behavior-related signals, which we have no reason to expect would target V1 alone). Although the model was allowed to assign weaker weight to LM inputs compared to V1 inputs if that improved the model's predictive performance, it was important to check that our conclusions held even in the absence of external input to LM. We, therefore, fitted models in which LM received no external input. Although these models provided a slightly worse statistical fit to the data (Extended Data Fig. 9a), they yielded the same conclusions regarding the presence of an approximate line attractor (Extended Data Fig. 9b,c).

Second, our models consisted of latent circuits with an equal number of excitatory and inhibitory units. A priori, we did not expect our results to be sensitive to the precise ratio of E–I units in the latent circuit, because a change in this ratio can be compensated for by changes in both recurrent inhibition and the mapping from latent units to observed neural activity. Nevertheless, to test this sensitivity directly, we fitted additional models with six E and two I units corresponding to the approximately 3:1 ratio of E–I neurons in the mouse cortex²⁶. These models also exhibited low-dimensional slow dynamics (Extended Data Fig. 9d–f), and their local currents were also dominated by inhibition, as in our main models (Fig. 3b, left), despite the higher E–I ratio (Extended Data Fig. 9g).

Interpretation of external inputs

Two of the external input channels in our data-driven models correspond to the two (go/no-go) visual stimuli. It is important to note that our go/no-go behavioral task does not allow for a clear separation of purely sensory from behavior-related (movement or reward) signals, as the behavioral report of decisions overlaps substantially with the stimulus presentation period. As such, external inputs inferred by our models should not be interpreted as purely visual inputs to the V1–LM circuit but as reflecting a combination of visual and other task-related signals (movement or reward-related signals). To reveal the contribution of such non-sensory signals to the inferred inputs, we compared these inputs in the correct go and no-go trials; given that the visual stimuli (orientation of the grating) associated with go and no-go categories were randomized across animals, any difference in average inferred input between the two conditions cannot be purely visual. These comparisons revealed that the models infer stronger inputs to the V1–LM circuit in go trials compared to no-go trials on average, especially after the main stimulus onset transient (Supplementary Fig. 3b), consistent with behavior-related variables providing extra input to V1 and LM late into the stimulus presentation. At the population level, these stronger inputs during go trials led to further movement along the approximate line attractor (Supplementary Fig. 3c). In fact, this was also true during false alarm trials, consistent with these late inputs being non-visual (Supplementary Fig. 3d).

Role of long-range connections in sustaining activity in the cortex

We found that long-range connections are crucial to sustaining activity in the visual cortex, whereas stimulus-specific external inputs are mainly responsible for onset and offset responses. This is consistent with recent recordings from the visual thalamus (dLGN²⁷), whose transient timecourse resembles our inferred external input. This prediction may, at first glance, appear to conflict with recent findings suggesting that silencing the thalamus via optogenetic activation of the thalamic reticular nucleus (TRN) leads to a rapid decay of activity in V1 (ref. 28). However, TRN activation inhibits not only dLGN but also higher-order thalamic areas (for example, pulvinar), which are thought to modulate cortico-cortical interactions^{29,30}. This manipulation could, therefore, effectively isolate V1 from other cortical areas, in which case the rapid

decay of cortical activity observed in Reinhold et al.²⁸ would be consistent with the fast time constants that we found in the isolated dynamics of our model's V1 population. More broadly, beyond visual networks, sustained cortical activity in decision-making or motor planning has also been shown to rely on multi-area interactions^{31,32}.

Brain state modulation of inter-area slow dynamics

Previous studies in primates suggested that disengaged or unattentive brain states are associated with reduced inter-area functional connectivity^{33,34}. Our inter-area consensus framework, therefore, predicts that the slow 'agree' mode would not contribute as much to neural dynamics in the disengaged states. To test this, we analyzed neural data from V1 and LM in disengaged trials, in which mice were grooming themselves instead of reacting to the stimulus. Autocorrelograms of V1–LM activity projected onto the agree and disagree modes (Extended Data Fig. 10a–c) reveal that it is only during engaged trials that activity is slower along the agree mode, relative to the disagree mode (this was true both for the unselective and selective agree/disagree pairs; Fig. 4c,i). Similarly, it is only during engaged trials that silencing one area causes activity to speed up in the other (Extended Data Fig. 10d–f).

Role of long-range connections in consensus building

We found that the coupled dynamics of V1 and LM implement a form of consensus algorithm, whereby the two areas progressively get to reconcile their views about the presence of a stimulus and its coarse orientation. The fairly simple nature of this consensus arguably reflects the simplicity of our experimental go/no-go task. However, we hypothesize that dynamic consensus is a general feature of cortical dynamics that could play out at finer scales and be modulated to meet complex behavioral demands. Notably, achieving fine-grained consensus requires detailed specificity in long-range connections between cortical areas. Such specificity could conceivably be achieved via trans-thalamic pathways, which appear to systematically mirror direct cortico-cortical pathways^{35,36}. Detailed gain modulation of thalamic neurons involved in these pathways (for example, pulvinar, known to send functionally specific projections to V1; ref. 37) could provide sufficient flexibility for regulating multiple modes of consensus between cortical areas. Indeed, Mo et al.³⁸ showed that inhibiting the trans-thalamic pathway between primary and higher-order somatosensory cortices in mice leads to a loss of learning-induced texture selectivity but no change in overall cell responsiveness to tactile stimuli. Our model of Fig. 6 would attribute such effects to a decrease in specific long-range connectivity affecting consensus in the selective mode useful for stimulus discrimination but not affecting the unselective mode useful for stimulus detection. More generally, richer forms of consensus arising from fine-grained connectivity could serve more complex computations—for example, the integration and reconciliation of bottom-up sensory information with top-down prior expectations³⁹. By integrating data from large-scale functional connectomics⁴⁰ with multi-area neural recordings during more complex tasks, our theoretical approach is ideally positioned to test such hypotheses and uncover the richer dynamics of brain-wide consensus.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-026-02437-3>.

References

1. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).

2. Douglas, R. J. & Martin, K. A. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27**, 419–451 (2004).
3. Markram, H. et al. Interneurons of the neocortical inhibitory system. *Nat. Rev. Neurosci.* **5**, 793–807 (2004).
4. Harris, K. D. & Shepherd, G. M. G. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
5. Apicella, A. & Marchionni, I. VIP-expressing GABAergic neurons: disinhibitory vs. inhibitory motif and its role in communication across neocortical areas. *Front. Cell. Neurosci.* **16**, 811484 (2022).
6. Cho, K. A., Shi, J., Phensy, A. J., Turner, M. L. & Sohal, V. S. Long-range inhibition synchronizes and updates prefrontal task activity. *Nature* **617**, 548–554 (2023).
7. Javadzadeh, M. & Hofer, S. B. Dynamic causal communication channels between neocortical areas. *Neuron* **110**, 2470–2483 (2022).
8. Schimel, M., Kao, T.-C., Jensen, K. T. & Hennequin, G. iLQR-VAE: control-based learning of input-driven dynamics with applications to neural data. In *Tenth International Conference on Learning Representations* <https://openreview.net/forum?id=wRODLHaAiW> (2022).
9. Rudy, B., Fishell, G., Lee, S. & Hjerling-Leffler, J. Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* **71**, 45–61 (2011).
10. Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D. & Ferster, D. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62**, 578–592 (2009).
11. Ahmadian, Y. & Miller, K. D. What is the dynamical regime of cerebral cortex? *Neuron* **109**, 3373–3391 (2021).
12. Sanzeni, A. et al. Inhibition stabilization is a widespread property of cortical networks. *eLife* **9**, e54875 (2020).
13. Ko, H. et al. Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87–91 (2011).
14. Haider, B., Häusser, M. & Carandini, M. Inhibition dominates sensory responses in the awake cortex. *Nature* **493**, 97–100 (2013).
15. Ganguli, S. et al. One-dimensional dynamics of attention and decision making in LIP. *Neuron* **58**, 15–25 (2008).
16. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
17. Nair, A. et al. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178–193 (2023).
18. Sylwestrak, E. L. et al. Cell-type-specific population dynamics of diverse reward computations. *Cell* **185**, 3568–3587 (2022).
19. Murphy, B. K. & Miller, K. D. Balanced amplification: a new mechanism of selective amplification of neural activity patterns. *Neuron* **61**, 635–648 (2009).
20. Ding, Z. et al. Functional connectomics reveals general wiring rule in mouse visual cortex. *Nature* **640**, 459–469 (2025).
21. Pandarinath, C. et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
22. Malonis, P. J., Vishnubhotla, A., Hatsopoulos, N. G., MacLean, J. N. & Kaufman, M. T. Combatting nonidentifiability to infer motor cortex inputs yields similar encoding of initial and corrective movements. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.18.464704> (2025).
23. Soldado-Magraner, J., Mante, V. & Sahani, M. Inferring context-dependent computations through linear approximations of prefrontal cortex dynamics. *Sci. Adv.* **10**, ead14743 (2024).
24. Sourmpis, C., Petersen, C. C. H., Gerstner, W. & Bellec, G. Biologically informed cortical models predict optogenetic perturbations. *eLife* **14**, RP106827 (2026).
25. Brenner, J. M., Beltramo, R., Gerfen, C. R., Ruediger, S. & Scanziani, M. A genetically defined tecto-thalamic pathway drives a system of superior-colliculus-dependent visual cortices. *Neuron* **111**, 2247–2257 (2023).
26. Rodarie, D., Verasztó, C. et al. A method to estimate the cellular composition of the mouse brain from heterogeneous datasets. *PLoS Comput. Biol.* **18**, e1010739 (2022).
27. Siegle, J. H. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
28. Reinhold, K., Lien, A. D. & Scanziani, M. Distinct recurrent versus afferent dynamics in cortical visual processing. *Nat. Neurosci.* **18**, 1789–1797 (2015).
29. Sherman, S. M. & Guillery, R. W. Distinct functions for direct and transthalamic corticocortical connections. *J. Neurophysiol.* **106**, 1068–1077 (2011).
30. Saalman, Y. B. & Kastner, S. Cognitive and perceptual functions of the visual thalamus. *Neuron* **71**, 209–223 (2011).
31. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532**, 459–464 (2016).
32. Guo, Z. V. et al. Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* **545**, 181–186 (2017).
33. Gregoriou, G. G., Gotts, S. J., Zhou, H. & Desimone, R. High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science* **324**, 1207–1210 (2009).
34. Bosman, C. A. et al. Attentional stimulus selection through selective synchronization between monkey visual areas. *Neuron* **75**, 875–888 (2012).
35. Halassa, M. M. & Sherman, S. M. Thalamocortical circuit motifs: a general framework. *Neuron* **103**, 762–770 (2019).
36. Shepherd, G. M. G. & Yamawaki, N. Untangling the cortico-thalamo-cortical loop: cellular pieces of a knotty circuit puzzle. *Nat. Rev. Neurosci.* **22**, 389–406 (2021).
37. Furutachi, S., Franklin, A. D., Aldea, A. M., Mrsic-Flogel, T. D. & Hofer, S. B. Cooperative thalamocortical circuit mechanism for sensory prediction errors. *Nature* **633**, 398–406 (2024).
38. Mo, C., McKinnon, C. & Sherman, S. M. A transthalamic pathway crucial for perception. *Nat. Commun.* **15**, 6300 (2024).
39. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
40. MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature* **640**, 435–447 (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Methods

Experimental procedures

No new experimental data were collected for the purposes of this study. The acquisition and preprocessing of data used in this study are described in detail in Javadzadeh and Hofer⁷. In brief, PV-Cre mice⁴¹ were used. Mice were of either sex and were 6–14 weeks old at the start of the experiment. Simultaneous electrophysiological recordings were obtained from retinotopically aligned sites in V1 and LM, in head-fixed stationary mice viewing visual stimuli on a display. Mice were trained on a go/no-go visual discrimination task, where they learned to distinguish between two static gratings oriented at +45° or −45°, with only one orientation being rewarded. The rewarded (go) stimulus was randomly assigned across animals. Mice indicated detection of the rewarded stimulus by licking a spout, which triggered reward delivery. Each grating was displayed for 500 ms. The inter-stimulus interval comprised a fixed 500-ms period followed by a variable delay drawn from a truncated exponential distribution (mean 4–6 seconds, truncated between 10 seconds and 30 seconds) to prevent trial predictability and excessively long intervals. To transiently suppress activity in one cortical area while recording from both, brief optogenetic silencing (150-ms duration) was applied during the presentation of grating stimuli. The onset of silencing varied in each trial, tiling the duration of the visual stimulus in 65-ms steps. Silencing was achieved by photo-activation of PV⁺ inhibitory interneurons expressing ChR2, following established approaches^{42,43}.

We subselected seven mice from Javadzadeh and Hofer⁷ for fitting data-driven models in this study, based on the criterion that the electrophysiological recordings contained at least one well-isolated single unit that was identified by the optogenetic perturbations as PV⁺. Models were fit using all trial types, but only trials in which the mice performed the task correctly were included in subsequent analyses, unless specified otherwise. The spiking activity of the recorded neurons was binned at 5-ms resolution and, for visualization, smoothed with a running average of 25 ms or five bins (Figs. 1b,c,g, 2a,d and 3c).

The model-free autocorrelation analyses of Figs. 4 and 5 and Extended Data Fig. 10 did not require isolation of single-unit PV⁺ neurons used for model validation; we could, therefore, include more mice ($n = 12$) for these analyses. Disengaged trials shown in Extended Data Fig. 10 are characterized by mice grooming themselves instead of reacting to the stimulus and are generally excluded for other analyses. The engaged trials include only correctly performed trials.

Latent circuit model of V1–LM data

Latent circuit dynamics. We modeled latent circuit dynamics as an input-driven recurrent neural network described by a standard firing rate equation⁴⁴. Specifically, the circuit's n -dimensional 'latent state' $\mathbf{z}$ evolved according to

$$\tau \dot{\mathbf{z}}(t) = -\mathbf{z}(t) + W\Phi(\mathbf{z}(t)) + B\mathbf{u}(t) \quad (3)$$

where $\tau = 20$ ms is a single-neuron characteristic time constant; W is a matrix of recurrent connectivity (see below); B is a matrix of input weights; and $\Phi(z) = \frac{1}{2}(z + \sqrt{z^2 + 0.1})$ is a soft rectified-linear activation function. (Note the presence of external inputs $\mathbf{u}(t)$ described in detail below.) The spiking activities of our N recorded neurons were then modeled as conditionally independent Poisson processes given the latent circuit's activity, $\mathbf{z}(t)$, with momentary firing rates $\mathbf{r}(t)$ given by:

$$\mathbf{r}(t) = \exp[\mathbf{C}\mathbf{z}(t) + \mathbf{d}] \quad (4)$$

$$\mathbf{y}(t)|\mathbf{z}(t) \sim \text{Poisson}(\mathbf{r}(t)). \quad (5)$$

Here, $\mathbf{C}$ is an $N \times n$ matrix of output weights and $\mathbf{d}$ is an N -dimensional vector of constant offsets. Equation (3) was discretized using a timestep $dt = 5$ ms. All model parameters were optimized to fit the

electrophysiological data (see below, Network training procedure section). Critically, W , $\mathbf{C}$ and B were constrained to reflect biophysical properties of the V1–LM network (see below; schematics in Fig. 1d–f).

Note that equation (3) does not include a constant input term. We found that including such a bias term caused the model to fall into local minima, consistently learning solutions with worse residual log-likelihoods (Extended Data Fig. 2e).

External inputs. Our model captures trial-by-trial variability in neural activity not only via the Poisson sampling step in equation (5) but also, and more importantly, through trial-by-trial fluctuations in the external inputs $\mathbf{u}(t)$. These (deterministically) produce variations in latent circuit activity according to equation (3) and, therefore, also in the neurons' firing rates (equation (4)). In the language of probabilistic modeling, the external inputs $\mathbf{u}$ constitute the model's latent variables.

Simultaneously inferring dynamics and external input is a fundamentally ill-posed problem, which our probabilistic model addresses by placing task-informed, non-stationary prior distributions on the latent inputs. Specifically, we used three input channels—that is, $\mathbf{u}(t) \equiv [u_0(t), u_1(t), u_2(t)]^T$ —each entering the latent circuit through input weights given by the corresponding column of the $n \times 3$ matrix B (equation (3)). For each input channel i , we assumed $u_i(t)$ to be (a priori) independently and normally distributed across timesteps—ensuring that any continuous/smooth fluctuations in firing rates could be accounted for only by recurrent dynamics in the latent circuit. Moreover, the variance of this Gaussian prior was given a channel-specific and trial-specific temporal profile reflecting the known timing of the corresponding stimulus:

$$u_i(t) \sim \mathcal{N}(0, \Sigma_0^i + \Sigma^i e_i(t)) \quad (6)$$

$$e_0(t) = 1 \text{ if laser on, } 0 \text{ otherwise} \quad (7)$$

$$e_1(t) = 1 \text{ if go stimulus on, } 0 \text{ otherwise} \quad (8)$$

$$e_2(t) = 1 \text{ if no-go stimulus on, } 0 \text{ otherwise} \quad (9)$$

where Σ_0^i and Σ^i are two positive variance parameters optimized alongside all other model parameters (see below).

Given that the laser input in our experiments had a direct effect only on inhibitory neurons, we constrained the first column of B (associated with $u_0(t)$) to be zero for all subpopulations except for the inhibitory neurons of the targeted area. Additionally, we ensured that the weights of this column of B were all positive. Finally, to eliminate the degeneracy that exists between the scale of the inputs $\mathbf{u}(t)$ (set by Σ_0^i and Σ^i as detailed above) and the scale of the matrix B , we constrained the norm of each column of B to be equal to $\sqrt{n/m}$ (where n is the number of units in the latent circuit and $m = 3$ is the number of input channels). The B matrices learned for all animals are shown in Extended Data Fig. 1.

Note that the posterior standard deviation of the inputs as shown in Fig. 1e, left, reflects the model uncertainty, whereas the trial-to-trial noise in the inputs is reflected in the standard deviation of the posterior mean (Fig. 1e, right; also see Supplementary Fig. 4).

Constraints on the latent circuit connectivity. We partitioned the latent circuit's activity $\mathbf{z}(t)$ into two halves, corresponding to the V1 and LM subcircuits, respectively (Fig. 1d). Within each subcircuit, we took the first half of the latent units to be excitatory and the other half to be inhibitory. This partitioning of the circuit into four subpopulations allowed us to enforce Dale's law, as well as the purely excitatory nature of long-range projections, by constraining the recurrent weight matrix W to have the following structure:

$$W = \begin{bmatrix} W_{EE}^{V1} & -W_{EI}^{V1} & W_{EE}^{LM \rightarrow V1} & 0 \\ W_{IE}^{V1} & -W_{II}^{V1} & W_{IE}^{LM \rightarrow V1} & 0 \\ W_{EE}^{V1 \rightarrow LM} & 0 & W_{EE}^{LM} & -W_{EI}^{LM} \\ W_{IE}^{V1 \rightarrow LM} & 0 & W_{IE}^{LM} & -W_{II}^{LM} \end{bmatrix} \quad (10)$$

with all elements of the various W blocks constrained to be positive. We enforced the sign constraints in our model by passing elements of W through a positive nonlinearity and multiplying W with a mask matrix containing the sign of each element. We note that, in related work, Jha et al.⁴⁵ proposed a method to learn linear latent dynamical systems constrained to follow Dale's law using a constrained quadratic optimization approach. Note that while W is structured such that only excitatory units project across areas, it is possible for the models to learn net inhibitory effects of one area on the activity of the other area, through $E \rightarrow I \rightarrow E$ pathways. See Supplementary Information, section S4 and Supplementary Fig. 5 for detailed analyses of the learned effective connectivity motifs between areas.

Structured sparsity constraint on the latents-to-neurons readout. The matrix C in equation (4), which determines how the firing rates of the recorded neurons (corresponding to rows) are assembled from the activity of the latent units (corresponding to columns), was constrained such that the neurons recorded in V1 (respective LM) would only be associated with V1 (respective LM) latent units. This was achieved by enforcing the following block structure (Fig. 1f):

$$C = \left[\begin{array}{cc|cc} \text{latent units} & & & \\ \hline s_E^{V1} C_E^{V1} & s_I^{V1} C_I^{V1} & 0 & 0 \\ 0 & 0 & s_E^{LM} C_E^{LM} & s_I^{LM} C_I^{LM} \end{array} \right] \begin{array}{l} \text{recorded} \\ \text{neurons} \end{array} \quad (11)$$

where each C_i is an element-wise positive matrix with unit-norm columns, and each corresponding s_i^c is a positive scalar. This per-block column-wise normalization of C balances the model internally by ensuring that all the latent units within each subpopulation have a similar effect on the activity of the observed neurons. Moreover, the inclusion of separate scale factors s_i^c allows the different E–I subpopulations to contribute to different degrees to the neural activity.

Notably, to facilitate interpretability of the latent circuit, we learned the model in such a way that it would unequivocally label each recorded neuron as being excitatory or inhibitory. We achieved this by including in the overall cost function (see below) a structured sparsity penalty on C that encourages each recorded neuron to be locally associated either with the excitatory latent units or with the inhibitory latent units but not with both types simultaneously. In other words, this penalty promotes parameter solutions in which the rows of C are non-zero either within the C_E^c block or within the C_I^c block (where c denotes the relevant cortical area). This penalty took the following form:

$$\mathcal{L}_{\text{sparsity}} = \lambda \sqrt{\sum_{n \in \text{neurons}} \| (C_E^c)_n \|^2 + \| (C_I^c)_n \|^2} \quad (12)$$

where $c_n \in \{V1, LM\}$ is the cortical area where neuron n was recorded; $(C^c)_n$ denotes the n -th row of the matrix block C^c ; and $\| \cdot \|$ denotes the L_2 norm. The scalar λ was set to 10^3 after a hyperparameter search.

Definition of putative excitatory and inhibitory cells. For models trained with the above constraints, we were able to assign each neuron a unique excitatory or inhibitory identity based on the learned readout matrix, C (Fig. 1f). For each neuron, we calculated the L_2 norms of the corresponding readout weights originating from the excitatory and inhibitory latent subpopulations separately and labeled the neuron as E or I according to which of the two norms was the largest.

Unconstrained models. For the ‘unconstrained models’, we imposed no specific circuit structure on W , which was allowed to be a full matrix with no sign constraints. For the matrix C , we did not impose the empty blocks shown in equation (11), such that there was no bias toward partitioning the latents into V1-related and LM-related groups. However, for better internal balancing and interpretability of the resulting latent dynamics, we did keep the column-wise norm constraints and associated scale parameters in equation (11) and applied the same regularization as in equation (12) to encourage grouping of the latents into similarly sized subpopulations, albeit with no E–I interpretation. See Extended Data Fig. 1a,b for a comparison of the learned parameters in unconstrained versus constrained models.

Network training procedure

Our latent circuit model, together with the prior distribution over external inputs and the Poisson observation noise model described above (equations 3–6), constitute a probabilistic generative model whose parameters we directly optimized to fit our spiking data. To this end, we used iLQR-VAE⁸, a generic control-based algorithm for learning probabilistic, input-driven latent dynamics from neural population recordings. iLQR-VAE learns model parameters θ that maximize a lower bound on the log-likelihood of the data, $\log p_\theta(\mathbf{y})$. This evidence lower bound (ELBO; ref. 46) is a standard objective, used when the true log-likelihood cannot be evaluated in closed form, as is the case in our model. The ELBO, denoted by $\mathcal{L}$, relies on an approximate posterior distribution over inputs, $q_\phi(\mathbf{u}|\mathbf{y})$:

$$\mathcal{L}(\mathbf{y}, \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{u}|\mathbf{y})} [\log p_\theta(\mathbf{y}|\mathbf{u}) + \log p_\theta(\mathbf{u}) - \log q_\phi(\mathbf{u}|\mathbf{y})] \leq \log p_\theta(\mathbf{y}). \quad (13)$$

In iLQR-VAE, $q_\phi(\mathbf{u}|\mathbf{y}) = \mathcal{N}(\mu_\theta(\mathbf{y}), \Sigma)$ is parametrized as a Gaussian distribution, whose mean $\mu_\theta(\mathbf{y})$ is defined as the most likely set of inputs given the data and the model parameters. This maximum a posteriori estimate can be efficiently obtained using the iLQR algorithm⁴⁷:

$$\mu_\theta(\mathbf{y}) = \arg \max_{\mathbf{u}} \sum_{t=1}^T \log p_\theta(\mathbf{y}_t|\mathbf{u}) + \log p_\theta(\mathbf{u}_t) \quad (14)$$

$$= \text{iLQRsolve}(\mathbf{y}, \theta) \quad (15)$$

As in Schimel et al.⁸, we defined the covariance Σ as a trial-independent, separable matrix—that is, as the Kronecker product of a spatial factor Σ_s and a temporal factor Σ_t , which were learned throughout training and shared across all training trials.

In summary, fitting our latent circuit model to the V1–LM spiking data involved jointly optimizing all model parameters θ and the approximate posterior parameters $\phi = \{\theta, \Sigma_s, \Sigma_t\}$ to minimize the following combined objective:

$$\mathcal{O}(\theta, \phi) = -\mathcal{L}(\mathbf{y}, \theta, \phi) + \mathcal{L}_{\text{sparsity}}(\theta) \quad (16)$$

Retraining models with a fraction of neurons. To establish that the posterior standard deviation indeed captures the uncertainty of the models, we analyzed how it changes when models have access to the activity of fewer recorded neurons. To do this, we took models that had been trained on the full set of neurons, and we froze all parameters except for the posterior covariance. We then trained the posterior covariance on datasets that were generated by sampling the original training set but only keeping a fraction of the neurons. Results shown in Supplementary Fig. 4a indicate that the posterior uncertainty increases as the number of observed neurons diminishes.

Log-likelihood computations

Computation of cross-validated log-likelihoods. To validate the performance of our model, we computed its ability to predict the

activity of held-out neurons, given firing rates inferred using the held-in neurons. We held out one neuron at a time. To predict the activity of held-out neuron j , we inferred inputs as $\hat{\mu}^k = \text{iLQRsolve}(\mathbf{y}_{-j}^k, \theta_{-j})$, where $\mathbf{y}_{-j}^k \in \mathbb{R}^{(N-1) \times T}$ is the spike trains of all neurons excluding neuron j in trial k , and θ_{-j} are the model parameters with the j -th row of $\mathbf{C}$ and $\mathbf{d}$ trimmed away. We then computed the predicted firing rates for all (both held-in and held-out) neurons r^k by unrolling the trajectories induced by the inputs $\hat{\mu}^k$ (using the full set of parameters θ). In turn, this allowed us to compute the log-likelihood of the spikes in trial k for the held-out neuron j , as

$$LL_j^k = \sum_t [\mathbf{y}_j^k(t) \log(\hat{r}_j^k(t) dt) - \hat{r}_j^k(t) dt - \log y_j(t)!]. \quad (17)$$

Computation of the empirical log-likelihood. As a baseline to compare the model predictions to, we computed the empirical log-likelihood for a trial k by evaluating the predicted activity for every neuron using that neuron's average activity across all the other trials from the same condition c , leading to a predicted firing rate timecourse

$$r_{j,\text{emp}}^k(t) = \frac{1}{N_c - 1} \sum_{\ell \in c, \ell \neq k} \mathbf{y}_j^\ell(t), \quad (18)$$

where N_c is the number of trials in condition c . Given these empirical firing rates, we computed the empirical log-likelihood for neuron j at trial k as

$$LL_{j,\text{emp}}^k = \sum_t [\mathbf{y}_j^k(t) \log(r_{j,\text{emp}}^k(t) dt) - r_{j,\text{emp}}^k(t) dt - \log y_j(t)!]. \quad (19)$$

Residual log-likelihood. We define the residual log-likelihood for a given neuron j as $LL_j^k - LL_{j,\text{emp}}^k$.

If this quantity is positive, it means that the prediction of the model for that neuron is more accurate than a prediction based on trial averaging—that is, that the model is able to capture meaningful single-trial variability in the data. Residual likelihoods were calculated separately for each neuron across 18 different conditions (two visual stimuli and nine silencing conditions for each visual stimulus) and then averaged across all trials and conditions.

Model selection

Choice of hyperparameters. To select the model hyperparameters n and m (number of latent state variables and input channels, respectively), we used a three-fold cross-validation approach. For each animal, we split the trials into three subsets. Then, for each possible pair of subsets among these three, we trained a model using the data from that pair and subsequently computed the held-out log-likelihoods on the remaining subset. Finally, we averaged the results over the three pairs, over animals and over neurons.

We first selected the optimal value of n for the model with three input channels ($m = 3$) corresponding to the visual and optogenetic stimuli, as described above. We explored model sizes ranging from $n = 8$ to $n = 24$ in increments of 4 and selected the minimal value of n after which the residual log-likelihood stopped improving (Extended Data Fig. 2a). Having selected and fixed the optimal value for n , we checked whether the choice $m = 3$ was optimal, using the same model selection procedure. When varying the number of input channels, we considered both (1) having multiple channels corresponding to each prior variance profile (see equations (7–9))—that is, multiple channels for each external stimulus (Extended Data Fig. 2b)—and (2) the addition of channels with temporally unmodulated prior variance (Extended Data Fig. 2c). Neither of those increased model performance relative to using $m = 3$, which is the minimal number of channels allowing to have one input per external stimulus. Note that models with additional input channels could, in theory, capture

timing difference in the visual input to V1 and to LM. However, we found that having one channel per input yielded the best performance on the validation fold.

Additionally, we compared our models to models with the same architecture but for which the inputs were not inferred and were, instead, fixed to follow the envelope corresponding to each external stimulus. This implied that their timecourse was constrained to be the same for every trial of a given condition. Those models performed considerably worse than the models with inferred inputs (Extended Data Fig. 2d). Other model hyperparameters, such as the spectral radius of $\mathbf{W}$ at initialization and the Adam learning rate, were fixed to values that allowed robust training. Final hyperparameter choices are reported in Table 1. All trials, irrespective of behavioral outcome, were included for log-likelihood calculation and model selection.

Selection of models for plotting and analysis. For the constrained models, having set the hyperparameters as described above, we trained 10 models with different random seeds (that is, different random initializations of the model) per animal. This was done to reduce the chance of getting stuck in local minima. Moreover, as our conclusions were dependent on the learned values of the long-range weights, and to avoid biasing our models, we varied the value of the long-range weights at initialization. More precisely, we varied the ratio of the norm of long-range weights to local weights at initialization between 1 and 1.6 in steps of 0.2. We discarded models that diverged during training (41 of 280 models in total). Of the remaining models, we then picked the best model for each animal, across initialization seeds and long-range weights, for further analyses and plotting. For each animal, the best model was selected by first subselecting the models that classified the known PV⁺ cells correctly as inhibitory (187 of 239—that is, 78.24% of the models; Extended Data Fig. 2f). Among these, we picked the model that yielded the highest cross-validated log-likelihood. Furthermore, we only included active cells (neurons whose spike count during the stimulus in control trials had a signal-to-noise ratio—that is, mean / s.d. over trials—larger than 1) for log-likelihood calculations.

For the unconstrained models, we used five random initialization seeds. However, as inhibitory cell identities were not defined in these models, we picked the best model based only on the held-out log-likelihood criterion explained above.

Alternative models

To assess the robustness of our findings, especially the existence of the approximate line attractor, on the specific model parametrization, we fitted several alternative models to the data. Results shown in Extended Data Fig. 9 indicate that the approximate line attractor persists under these model variations.

Models with no external inputs to LM. In our main models, visual inputs target both V1 and LM. We additionally included alternative models in which the visual inputs were restricted to target V1 only. This was done by masking part of the $\mathbf{B}$ matrix, such that the two visual input channels had zero weights onto the LM latents.

Models with a 3:1 ratio of excitatory to inhibitory latent units. Our main models consist of latent circuits with equal numbers of excitatory and inhibitory units. To test the sensitivity of results to the latent E–I ratio, we fitted additional models with six E and two I units corresponding to the approximately 3:1 ratio of E–I neurons in the mouse cortex²⁶. We initialized these models in a way that preserved the relative strength of inhibition and excitation relative to our original circuit models.

Calculating covariances

In Extended Data Fig. 3a,b, we calculated $N \times N$ noise covariance matrices in both data and model-predicted activity as

$$\Sigma = \frac{1}{T \sum_c K_c} \sum_{c=1}^C \sum_{k=1}^{K_c} \sum_{t=1}^T (\mathbf{y}_{t,c}^k - \bar{\mathbf{y}}_{t,c}) (\mathbf{y}_{t,c}^k - \bar{\mathbf{y}}_{t,c})^T \quad (20)$$

where c indexes conditions (two visual stimuli and nine silencing conditions per stimulus), and $\mathbf{y}_{t,c}^k$ is an $N \times 1$ vector denoting spike count of N neurons in 25-ms bins, in control condition c (no optogenetic stimulation), trial k and time t (K_c : number of trials in condition c ; T : number of timepoints, N : number of neurons). $\bar{\mathbf{y}}_{t,c}$ is the trial-averaged activity in condition c . For calculating model covariances, we sampled pseudo-observations $\mathbf{y}$ from a Poisson distribution whose mean was taken to be the posterior predicted firing rates. All trials, irrespective of behavioral outcome, were used for calculating covariances. Variances in Extended Data Fig. 3a,b are the diagonal values of Σ , and cross-covariances are its off-diagonal values.

Linearization of the dynamics

Around a (approximate) fixed point $\mathbf{z}_f$, the dynamics in equation (3) can be Taylor-expanded to first order, leading to a linear dynamical system whose dynamics matrix is given by the Jacobian $\mathbf{A}$:

$$\mathbf{A} = -\mathbf{I} + \frac{\mathbf{W}\Phi'(\mathbf{z}_f)}{\mathbf{W}_{\text{eff}}} \quad (21)$$

Here, $\mathbf{W}_{\text{eff}}$ can be thought of as a matrix of ‘effective connectivity’. For a given trial k , we defined $\bar{\mathbf{z}}^k$ as the time-averaged activity either before or during stimulus—that is, $\bar{\mathbf{z}}^k = \frac{1}{\Delta} \sum_{t=T_0}^{T_0+\Delta} \mathbf{z}_t^k$ with $\Delta = 400$ ms, $T_0 = -400$ ms for the pre-stimulus window and $T_0 = 100$ ms for the stimulus window (T_0 is measured relative to visual stimulus onset). This choice was motivated by the fact that the dynamics exhibited very small velocities in these time windows (Fig. 3e). We then defined $\mathbf{z}_f = \frac{1}{K} \sum_k \bar{\mathbf{z}}^k$.

Computation of the dynamics distance

In Fig. 2c, we computed the similarity between the dynamics of the model for different animals, as a normalized Procrustes distance (see refs. 48 and 49) between their linearized dynamics—that is, as

$$d(A_i, A_j) = \min_{U \in O(n)} 1 - \frac{\text{Tr} [A_i^T (U A_j U^T)]}{\|A_j\|_F \|A_i\|_F}, \quad (22)$$

where A_i and A_j denote the linearized learned dynamics for animals i and j (obtained as described in the Linearization of the dynamics section); $\|\cdot\|_F$ denotes the Frobenius norm; and U is an orthogonal (rotation) matrix (optimization over U is necessary in order to account for the fact that the dynamics may be equivalent up to a rotation). We used the average distance $d(A_i, A_j)$ across all pairs of animals (that is, all (i, j) such that $i > j$), as our measure of consistency of the learned dynamics across animals.

As shown in Ostrow et al.⁴⁹, $d(\cdot, \cdot)$ is a valid distance metric, bounded between 0 and 1, which computes the similarity of the vector fields of two dynamical systems. Whereas Ostrow et al.⁴⁹ applied this analysis to dynamical systems identified via delay embedding of the dynamics, we instead apply it directly to the linearized dynamics of our model.

To perform the minimization in equation (22), we parametrized the orthogonal matrix U using a Cayley transformation⁴⁹. We initialized the optimization using an analytical expression that is guaranteed to yield the optimal unitary transformation U_{ij}^* , in the case where A_i and A_j are indeed exactly related by a unitary transformation. As pointed out in Ostrow et al.⁴⁹, the optimization landscape is disjoint for U matrices with $\det U = 1$ and $\det U = -1$. Thus, for each pair of dynamics matrices (A_i, A_j) , we performed two independent optimization runs: one starting from U_{ij}^* and another one starting from a modified version of U_{ij}^* that flipped its determinant (simple row permutation). We then retained the best of the two solutions.

Comparing the model’s ability to capture the effect of optogenetic perturbations

To evaluate how well the models captured the effect of simulated optogenetic perturbations (Fig. 2d), we first evaluated the inferred input for each no-go, no-laser trial. We then ran the model dynamics forward with those inputs while additionally perturbing the inhibitory model population in either V1 and LM, depending on which area expressed ChR2 in our experiments (different across animals). Note that these perturbations were simulated for each trial, based on the inferred input for that trial. We could then compare the neural responses predicted by the latent circuit model to the corresponding photo-stimulation responses observed in the experiments. Specifically, we used a simulated pulse of optogenetic input modeled as

$$p(t) = \begin{cases} 1 & \text{if } t \in [t_{\text{laser}}, t_{\text{laser}} + \Delta] \\ 0 & \text{otherwise} \end{cases} \quad (23)$$

where t_{laser} is the onset time of laser stimulation in the relevant silencing condition in the experiments and $\Delta = 150$ ms is the laser duration. We assumed that this input influenced the latent units via a weight vector $\mathbf{B}_p$, whose elements were non-zero only for the inhibitory latent units of the stimulated area. We optimized the non-zero elements of $\mathbf{B}_p$ to maximize the log-likelihood for the spike trains of the known PV⁺ cells in the relevant perturbation trials. We then measured the average predicted perturbation-induced change (relative to no-perturbation), $\Delta \hat{\mathbf{r}} = \hat{\mathbf{r}}_{\text{control}} - \hat{\mathbf{r}}_{\text{perturbation}}$ in the rest of the neurons during the stimulation time window and compared it to the same quantity, $\Delta \mathbf{r}$, measured in the data. We report the quality of fit as the Pearson’s correlation between $\Delta \hat{\mathbf{r}}$ and $\Delta \mathbf{r}$. This is plotted for one animal in Fig. 2d, middle, and for the rest of the animals in Extended Data Fig. 3c.

As a comparison (Fig. 2d), we repeated the above for ‘control-only’ models that were trained on control trials without optogenetic perturbation (two-thirds of the control trials were used for training). We trained a minimum of 12 models per animal and chose the best model following the same procedure used for the default models (see Selection of models for plotting and analysis section). (For one of the animals, we performed 30 fits instead of 12, and none resulted in correct classification of all PV⁺ neurons; for that animal, we only used the log-likelihood criterion for model selection.)

Spike width histograms

We extracted the average spike waveforms for each neuron, and the spike width was defined as the width of this waveform at 10% of its full amplitude (Fig. 2e).

Analysis of the role of inhibition in the dynamics

To evaluate the role of inhibition in stabilizing the dynamics (Fig. 2f), we measured the stability of our latent circuit dynamics, in the presence or absence of inhibition. We measured stability before and during stimulus presentation by computing the effective connectivity $\mathbf{W}_{\text{eff}}$ (see equation (21) in the Linearization of the dynamics section). We then computed the largest real part of the eigenvalues of the effective linear dynamical system, $\lambda_{\text{max}} = \max_i (\Re(\lambda_i))$, where λ_i are the eigenvalues of $\mathbf{W}_{\text{eff}}$. A (linearized) network is said to be ‘inhibition stabilized’ if $\lambda_{\text{max}} < 1$ (stable) when computed on the full $\mathbf{W}_{\text{eff}}$ but $\lambda_{\text{max}} > 1$ (unstable) when all the inhibitory weights in $\mathbf{W}_{\text{eff}}$ are set to zero.

Connectivity strength as a function of the noise correlation

In Fig. 2g, we computed the noise correlation matrix of the mean-subtracted latent circuit responses of the V1 excitatory subcircuit during control no-go trials, $\mathbf{z}_t$, as follows:

$$\Sigma = D^{-1/2} \left[\frac{1}{KT} \sum_{k=1}^K \sum_{t=1}^T (\mathbf{z}_{t,k} - \bar{\mathbf{z}}_t) (\mathbf{z}_{t,k} - \bar{\mathbf{z}}_t)^T \right] D^{-1/2} \quad (24)$$

where t and k denote time bin and trial, respectively; $\bar{\mathbf{z}}_t = \frac{1}{K} \sum_{k=1}^K \mathbf{z}_{t,k}$; and D is a diagonal matrix of single-neuron variances—that is, $D_{ii} = \frac{1}{KT} \sum_{k=1}^K \sum_{t=1}^T (z_{t,k}^i - \bar{z}_t^i)^2$.

In Fig. 2g, we plot the effective connection weights, $W_{\text{eff}}^{i,j}$ (computed based on the stimulus period as described by equation (21)), for pairs of excitatory latent units ($i < j$), as a function of the pairs' noise correlation $\Sigma^{i,j}$ (more specifically, we binned pairs based on their noise correlation and created box-and-whisker plots of the effective connection weights for each bin).

When comparing with the circuit at initialization (Fig. 2g, top), we considered W_{eff} and $\mathbf{z}_t$ to be the connectivity matrix and latent trajectories at the first iteration of training the model. We repeated the same procedure for go trials, with similar results (Extended Data Fig. 3d).

Calculating recurrent and external currents

For analyses described in Fig. 3a,b, we defined external and recurrent currents as $\text{ext}(t) = \sum_{i=1}^n [\mathbf{B}\mathbf{u}(t)]_i$ and $\text{rec}(t) = \sum_{i=1}^n [\mathbf{W}\Phi(\mathbf{z}(t))]_i$, respectively.

Sensitivity of the networks

In Fig. 3a,b,j, we evaluated the sensitivity of the latent circuit to changes in the inputs versus changes in the recurrent weights by running the network dynamics forward, using the inputs inferred from the data for every test trial, but including a gain γ that we used to scale down either the input matrix $\mathbf{B}$ (equation (25)) or the connectivity matrix $\mathbf{W}$ (equation (27)):

$$\tau \dot{\mathbf{z}}^{y_u}(t) = -\mathbf{z}^{y_u}(t) + \mathbf{W}\Phi(\mathbf{z}^{y_u}(t)) + \gamma \mathbf{B}\mathbf{u}(t) \quad (25)$$

$$\mathbf{r}^{y_u} = \exp(\mathbf{C}\mathbf{z}^{y_u} + \mathbf{d}) \quad (26)$$

versus

$$\tau \dot{\mathbf{z}}^{y_w} = -\mathbf{z}^{y_w}(t) + \gamma \mathbf{W}\Phi(\mathbf{z}^{y_w}(t)) + \mathbf{B}\mathbf{u}(t) \quad (27)$$

$$\mathbf{r}^{y_w} = \exp(\mathbf{C}\mathbf{z}^{y_w} + \mathbf{d}), \quad (28)$$

We computed the sensitivity by measuring changes in the total activity in no-go trials, either before or during stimulus onset, and normalizing those to the activity obtained for $\gamma = 1$ —that is, $S = \sum_{t=t_1}^{t_2} \mathbf{r}^y(t) / \sum_{t=t_1}^{t_2} \mathbf{r}(t)$ where $\mathbf{r}(t) = \mathbf{r}(t) - \mathbf{r}_{bs}$, with $\mathbf{r}_{bs}$ the average baseline (pre-stimulus) activity.

We used the same approach to compute the sensitivity separately to either local or long-range weights, which was done by applying the gain to the corresponding local (W^{LM} and W^{V1}) or long-range ($W^{LM \rightarrow V1}$ and $W^{V1 \rightarrow LM}$) blocks of the $\mathbf{W}$ matrix (equation (10)).

Intrinsic flow and velocity

In Fig. 3d, we plot the velocity field of the intrinsic dynamics (that is, dynamics in the absence of external inputs), projected onto the subspace spanned by the top two principal components of the latent trajectories. Note that projections onto the principal components were used only for visualization purposes, and all analyses were performed using the full-dimensional dynamics. We first performed a singular value decomposition on the trial-averaged latent activity in no-go trials $\mathbf{Z} \in \mathbb{R}^{n \times T}$ as $\mathbf{Z} = \mathbf{U}\Sigma\mathbf{V}^T$, before defining $\tilde{\mathbf{U}} = [\mathbf{U}_1, \mathbf{U}_2] \in \mathbb{R}^{n \times 2}$ as the top two principal components. We then computed the projected velocity field at each point in the two-dimensional space, $\mathbf{x} = (x, y)$, as $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^2$, where

$$\mathbf{v}(\mathbf{x}) = \tilde{\mathbf{U}}^T \dot{\mathbf{z}}(\tilde{\mathbf{U}}\mathbf{x}) \quad (29)$$

and the function $\dot{\mathbf{z}}(\cdot)$ is given by

$$\tau \dot{\mathbf{z}}(\xi) = -\xi + \mathbf{W}\Phi(\xi). \quad (30)$$

To compute the velocity in Fig. 3e, we similarly used equation (30), but we used the no-go trial-averaged latent trajectories (without projecting

onto principal components) for ξ . In Fig. 3h, we followed the same procedure but using the $\mathbf{Z}$ and $\mathbf{W}$ restricted to each area. In this case, the two principal components were similarly extracted from the area-restricted latents. In Fig. 3d and Fig. 3h, the magnitude of the velocity indicated by the color bar corresponds to the norm of the n -dimensional velocity vector and not the projected one.

Network time constants and line attractor score

For analyses in Fig. 3f,g,i,j, we linearized the dynamics around the average value of the latents across trials and time, either during or before the stimulus, and computed the eigenvalues and eigenmodes of the linearized dynamics $\mathbf{A}$ (see Linearization of the dynamics section). In this continuous-time linear dynamical system, each eigenmode j evolves in time according to $e^{\lambda_j t}$, where τ is the single-neuron time constant and λ_j is the eigenvalue of mode j . The characteristic decay timescale of each mode is given by $\tau_j = \frac{\tau}{|\text{Re}(\lambda_j)|}$. Assuming the modes are ordered such that $0 > \text{Re}(\lambda_0) \geq \dots \geq \text{Re}(\lambda_n)$ —that is, $\tau_0 \geq \dots \geq \tau_n - \tau_0$ defines the slowest timescale in the dynamics.

To calculate the time constants in the V1-only or LM-only networks in Fig. 3i, we followed the same procedure but used $\mathbf{W}$ and $\mathbf{Z}$ restricted to each individual area to compute the linearized dynamics. When comparing the time constants of these single-area networks to the full network, in order to control for their smaller size, we constructed subnetworks of the size of each individual area (that is, of size $n/2$), sampled randomly from the full latent network (500 random subsets and excluding any subselection that would correspond to the V1 or LM network).

To quantify the existence of a line attractor in the dynamics, we compute the 'line attractor score', defined as in Nair et al.¹⁷ as a log ratio of the slowest to the second-slowest time constant of the network dynamics—that is, $\log(\tau_0/\tau_1)/\log 2$. A true line attractor would correspond to an infinite line attractor score. A score of 1 means that the slowest mode is twice as slow as the next mode. A score of 0 means that the first two slowest modes have the same time constant (as happens, for example, when these two modes define a plane with rotational dynamics—that is, the imaginary parts of their eigenvalues are non-zero).

Minimal E-I networks

Our minimal E-I networks (Figs. 4 and 6) are described as linear rate models, consisting of two areas, where each area's connectivity is given by

$$\mathbf{W}_{\text{local}} = \begin{bmatrix} e & -i \\ e & -i \end{bmatrix} \quad (31)$$

where e and i are the strength of excitatory and inhibitory connections, respectively. The activity in the full network evolves as

$$\tau \dot{\mathbf{r}} = -\mathbf{r} + \mathbf{W}\mathbf{r} + \mathbf{u}(t) \quad (32)$$

where $\mathbf{u}(t)$ is an external input, which is zero unless otherwise specified,

$$\mathbf{r} = \begin{bmatrix} r_{V1}^E \\ r_{V1}^I \\ r_{LM}^E \\ r_{LM}^I \end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} e & -i & \ell & 0 \\ e & -i & \ell & 0 \\ \ell & 0 & e & -i \\ \ell & 0 & e & -i \end{bmatrix} \quad (33)$$

and ℓ is the strength of long-range excitatory connections. This is the minimal network architecture depicted in Fig. 4b.

We can show that the orthonormal basis $\mathbf{Q}$ consisting of vectors $\mathbf{Q} = [\mathbf{b}_a, \mathbf{u}_a, \mathbf{b}_d, \mathbf{u}_d]$ ('b' for 'balanced', 'u' for 'unbalanced', 'a' for 'agree', 'd' for 'disagree'), where

$$\mathbf{b}_a = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{u}_a = \frac{1}{2} \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \quad \mathbf{b}_d = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \quad \mathbf{u}_d = \frac{1}{2} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \quad (34)$$

is a Schur basis of W such that $\tilde{W} = Q^T W Q$ is the upper triangular matrix

$$\tilde{W} = \begin{bmatrix} e - i + \ell & e + i + \ell & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & e - i - \ell & e + i - \ell \\ 0 & 0 & 0 & 0 \end{bmatrix}. \quad (35)$$

This upper triangular form describes feedforward connectivity in the new basis Q and reveals the existence of two separate functional subnetworks, respectively describing the dynamics of agreement and disagreement between V1 and LM. The dynamics of each functional subnetwork are characterized by balanced amplification due to the feedforward weight from the unbalanced to the balanced mode¹⁹. See Supplementary Information, section S5.1 for more details, including the interpretation of different elements of the Schur form.

We also considered a version of the above minimal model that also incorporated a notion of selectivity for go versus no-go stimuli (Fig. 4c). Specifically, we split every E and I population in V1 and LM into two subpopulations, one receiving direct go input and the other receiving direct no-go input. This resulted in a circuit with eight units, with recurrent connectivity parameterized as

$$W = \begin{bmatrix} e + e_s & e & -(i + i_s) & -i & \ell + \ell_s & \ell & 0 & 0 \\ e & e + e_s & -i & -(i + i_s) & \ell & \ell + \ell_s & 0 & 0 \\ e + e_s & e & -(i + i_s) & -i & \ell + \ell_s & \ell & 0 & 0 \\ e & e + e_s & -i & -(i + i_s) & \ell & \ell + \ell_s & 0 & 0 \\ \ell + \ell_s & \ell & 0 & 0 & e + e_s & e & -(i + i_s) & -i \\ \ell & \ell + \ell_s & 0 & 0 & e & e + e_s & -i & -(i + i_s) \\ \ell + \ell_s & \ell & 0 & 0 & e + e_s & e & -(i + i_s) & -i \\ \ell & \ell + \ell_s & 0 & 0 & e & e + e_s & -i & -(i + i_s) \end{bmatrix} \quad (36)$$

where e_s , i_s and ℓ_s control the stimulus selectivity of local excitatory and inhibitory and long-range excitatory connections, respectively. The Schur decomposition of this network, along with its interpretation in terms of time constants, can be found in Supplementary Information, section S5.

Autocorrelation of neural data

The autocorrelation of the agree (a) and disagree (d) neural activity patterns across V1 and LM are defined, respectively, as the sum and difference of the average empirical spike counts binned at 5 ms, $s(t)$, within each area—that is,

$$a_k(t) = \frac{1}{N_{V1}} \sum_{i \in V1} s_k^i(t) + \frac{1}{N_{LM}} \sum_{i \in LM} s_k^i(t) \quad (37)$$

$$d_k(t) = \frac{1}{N_{V1}} \sum_{i \in V1} s_k^i(t) - \frac{1}{N_{LM}} \sum_{i \in LM} s_k^i(t) \quad (38)$$

$$\tilde{a}_k(t) = a_k(t) - \frac{1}{K} \sum_{k' \in K} a_{k'}(t) \quad (39)$$

$$\tilde{d}_k(t) = d_k(t) - \frac{1}{K} \sum_{k' \in K} d_{k'}(t). \quad (40)$$

We define the autocorrelation of the agree mode as the autocovariance normalized to the overall variance:

$$C_k^a(\tau) = \frac{\langle \tilde{a}_k(t) \tilde{a}_k(t + \tau) \rangle_t}{\langle \tilde{a}_k(t) \tilde{a}_k(t) \rangle_t} \quad (41)$$

where $\langle \cdot \rangle_t$ denotes an average over time bins t that are such that both t and $t + \tau$ fall within the relevant time window. This time window was $[-400, 0]$ ms ('pre') or $[100, 500]$ ms ('during') relative to stimulus onset. The autocorrelation of the disagree mode, $C_k^d(\tau)$, is defined analogously. See Extended Data Fig. 7b for a distribution of marginal variances (denominator in equation (41)) in the agree and disagree modes.

In Fig. 4h,i, we report the mean autocorrelation and its standard error across all correct control go and no-go trials and all animals. Note that mean subtraction in equation (39) was done separately per animal/condition for go and no-go trials.

Our minimal models of V1–LM dynamics (Fig. 4b,c) also make predictions for the decay timescales of balanced versus unbalanced modes. Estimating the autocorrelation time constant of these modes required estimating the E–I identity of each recorded neuron. For this, we used the identities inferred by the latent circuit models and computed the momentary contributions of the balanced and unbalanced agree (a_b , a_u) or disagree (d_b , d_u) modes to the recorded activity as

$$a_b(t) = \frac{1}{n_{V1^E}} \sum_{i \in V1^E} s_k^i(t) + \frac{1}{n_{V1^I}} \sum_{i \in V1^I} s_k^i(t) + \frac{1}{n_{LM^E}} \sum_{i \in LM^E} s_k^i(t) + \frac{1}{n_{LM^I}} \sum_{i \in LM^I} s_k^i(t) \quad (42)$$

$$a_u(t) = \frac{1}{n_{V1^E}} \sum_{i \in V1^E} s_k^i(t) - \frac{1}{n_{V1^I}} \sum_{i \in V1^I} s_k^i(t) + \frac{1}{n_{LM^E}} \sum_{i \in LM^E} s_k^i(t) - \frac{1}{n_{LM^I}} \sum_{i \in LM^I} s_k^i(t) \quad (43)$$

$$d_b(t) = \frac{1}{n_{V1^E}} \sum_{i \in V1^E} s_k^i(t) + \frac{1}{n_{V1^I}} \sum_{i \in V1^I} s_k^i(t) - \frac{1}{n_{LM^E}} \sum_{i \in LM^E} s_k^i(t) - \frac{1}{n_{LM^I}} \sum_{i \in LM^I} s_k^i(t) \quad (44)$$

$$d_u(t) = \frac{1}{n_{V1^E}} \sum_{i \in V1^E} s_k^i(t) - \frac{1}{n_{V1^I}} \sum_{i \in V1^I} s_k^i(t) - \frac{1}{n_{LM^E}} \sum_{i \in LM^E} s_k^i(t) + \frac{1}{n_{LM^I}} \sum_{i \in LM^I} s_k^i(t) \quad (45)$$

Having defined these projections, we follow the same procedure as in equations (39)–(41) to calculate autocorrelations. Results are shown in Extended Data Fig. 7a.

Quantifying autocorrelation differences

To quantify the difference between the autocorrelation functions of neural activity projected onto the agree versus disagree modes (Fig. 4h,i), we first identified the time lag at which the average agree/disagree autocorrelation function reached its maximum ($t_{\max}$). We then quantified the difference, denoted by $\Delta_{\max}$, between the agree and disagree autocorrelation functions at this timepoint (see the caption of Fig. 4). In most autocorrelation plots, the first timepoint (5 ms) exhibited refractoriness and was excluded, with plots beginning at a time lag of 10 ms. Similarly, the 5-ms point was omitted when quantifying $t_{\max}$ and $\Delta_{\max}$.

Projected autocorrelations for selective/unselective modes

In Fig. 4i, to estimate the timecourse of the selective and unselective modes from the neural data, we computed two indices for each neuron. The first index measured 'unselective' responsiveness—that is, how much more each neuron responded to either stimuli (go/no-go), relative to baseline. The second index measured 'selective' responsiveness—that is, how much each neuron preferred the go stimulus over the no-go stimulus. This resulted in two vectors of indices:

$$\mathbf{w}^{\text{unsel}} = \frac{1}{TK} \sum_k \sum_{c=\text{go}, \text{no-go}} \left(\sum_{t=0}^{500} \mathbf{s}_k^c(t) - \sum_{t=-500}^0 \mathbf{s}_k^c(t) \right) \quad (46)$$

$$\mathbf{w}^{\text{sel}} = \frac{1}{TK} \sum_k \left(\sum_{t=0}^{40} \mathbf{s}_k^{\text{go}}(t) - \sum_{t=0}^{40} \mathbf{s}_k^{\text{no-go}}(t) \right) \quad (47)$$

We then used these indices to compute weighted averages of the neural activity at each timestep, $\alpha_k^{\text{unsel}}(t) = \mathbf{w}^{\text{unsel}T} \mathbf{s}_k(t)$ and $\alpha_k^{\text{sel}}(t) = \mathbf{w}^{\text{sel}T} \mathbf{s}_k(t)$.

The distributions of unselective and selective weights ($\mathbf{w}^{\text{unsel}}$ and $\mathbf{w}^{\text{sel}}$) across V1 and LM neurons are shown in Extended Data Fig. 7d,e, and their relationship with one another is shown in Extended Data Fig. 7c. The elements of $\mathbf{w}^{\text{unsel}}$ were biased toward positive values (Extended Data Fig. 7d), as most recorded neurons responded to visual stimuli by increasing their firing rates. For $\mathbf{w}^{\text{sel}}$, by contrast, the distribution was symmetric. This is because we measured responses early during stimulus presentation—that is, likely before any go-stimulus-related behavior could break the symmetry in the neural responses to the two stimuli (Extended Data Fig. 7e). The choice of a small time window to compute $\mathbf{w}^{\text{sel}}$ also ensured that the selective index for a neuron was not corrupted by its unselective stimulus responsiveness—that is, that $\mathbf{w}^{\text{sel}}$ was not directly correlated with $\mathbf{w}^{\text{unsel}}$ (Extended Data Fig. 7c). This was important to establish that the slow time constant of the autocorrelation in agree-selective mode was not simply due to the correlation of this mode with the agree-unselective one but, instead, depended on the stimulus selectivity of neurons.

Alignment of the latent circuits' eigenmodes onto agree unselective/selective modes

For our latent circuit models, we could ask whether their two agree modes (unselective and selective) bore any relationship with their two slowest eigenmodes. Eigenmodes were computed for the Jacobian of the latent circuit dynamics linearized around the average activity in no-go trials during stimulus presentation (see Linearization of the dynamics section) and were sorted from slowest to fastest according to their associated eigenvalues. Similarly, we could estimate the latent circuit's unselective and selective agree modes by computing, for each latent unit, similar indices of unselective/selective responsiveness as we had computed for recorded neurons in Fig. 4i (see equations 46 and 47). For the selective indices, activity of the latent circuit during the onset period (0–100 ms from the stimulus presentation) was used, as the latent circuit activity was strongly driven by external inputs during this time window (Fig. 3a). This yielded two normalized vectors, whose overlaps with the eigenvectors we evaluated, by calculating the absolute value of their dot product. These overlaps are shown in Extended Data Fig. 7f.

Details of Fig. 6

In Fig. 6, we simulated the linear dynamics of the minimal circuit model of Fig. 4c—that is, equations 32 and 36—with parameters $\tau = 10$ ms, $e = 2$, $i = 2$, $e_s = 0$, $i_s = 0$, $\ell = 0$ and ℓ_s taking values in the set $\{0, 0.5, 0.9\}$. The input to the network was $\mathbf{u}(t) = \alpha(t/\tau') \mathbf{u}_0$ with $\tau' = 15$ ms—that is, it was the product of a scalar temporal envelope (Fig. 6a, bottom) $\alpha(t) = [t^3 e^{-t} + \frac{1}{2}(1 - e^{-t})]H(t - t_{\text{stim}})$ (where $H(\cdot)$ denotes the Heaviside function) and a spatial input pattern $\mathbf{u}_0$, which expressed how much each subpopulation was driven by the 'visual' stimulus. For Fig. 6b–d, we set $\mathbf{u}_0 = (1, 0, 1, 0, -0.6, 0, -0.6, 0)^T$, whereas, for Fig. 6e–g, we set $\mathbf{u}_0 = (1, 0, 1, 0, 0, 0.6, 0, 0.6)^T$ (see gray insets).

Statistics and reproducibility

Statistical tests are indicated in the corresponding figure legends. Exact n values are provided in the figure legends and Methods. We used two-sided Wilcoxon rank-sum tests for independent group

comparisons and two-sided Wilcoxon signed-rank tests for paired comparisons, unless otherwise stated. No statistical method was used to predetermine sample size; sample sizes were based on previous literature and what is conventional for the field. No data were excluded from the analyses unless explicitly stated otherwise. The analyses were not randomized, and investigators were not blinded during analysis and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Fitted iLQR-VAE models used for the model-based analyses, including the training and test datasets, are available on figshare at <https://doi.org/10.6084/m9.figshare.31142140> (ref. 50). Preprocessed experimental data are available on figshare at <https://doi.org/10.6084/m9.figshare.30983044> (ref. 51).

Code availability

Code used to fit the iLQR-VAE models is available at https://github.com/marineschimmel/v1lm_nn. The repository includes model-fitting code and instructions for running the model-fitting pipeline as well as some example analyses. Fitted models used in the paper are available on figshare at <https://doi.org/10.6084/m9.figshare.31142140> (ref. 50).

References

- Hippenmeyer, S. et al. A developmental switch in the response of DRG neurons to ETS transcription factor signaling. *Nature* **435**, 88–93 (2005).
- Cardin, J. A. et al. Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* **459**, 663–667 (2009).
- Guo, Z. V. et al. Flow of cortical activity underlying a tactile decision in mice. *Neuron* **81**, 179–194 (2014).
- Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, 2005).
- Jha, A., Gupta, D., Brody, C. D. & Pillow, J. W. Disentangling the roles of distinct cell classes with cell-type dynamical systems. *Adv. Neural Inf. Process. Syst.* **37**, 33668–33690 (2024).
- Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations (ICLR, 2014)*.
- Li, W. & Todorov, E. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *Proceedings of the First International Conference on Informatics in Control, Automation and Robotics* (eds Araújo, H., Vieira, A., Braz, J., Encarnação, B. & Carvalho, M.) 222–229 <https://doi.org/10.5220/000114390220229> (SciTePress, 2004).
- Williams, A. H., Kunz, E., Kornblith, S. & Linderman, S. Generalized shape metrics on neural representations. *Adv. Neural Inf. Process. Syst.* **34**, 4738–4750 (2021).
- Ostrow, M., Eisen, A. J., Kozachkov, L. & Fiete, I. R. Beyond geometry: comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In *Proc. 37th Conference on Neural Information Processing Systems (NeurIPS, 2023)*.
- Javadzadeh, M. Fitted models. *Figshare* <https://doi.org/10.6084/m9.figshare.31142140.v1> (2026).
- Javadzadeh, M. Simultaneously recorded neural activity across two visual cortical areas in mice (V1 and LM). *Figshare* <https://doi.org/10.6084/m9.figshare.30983044.v1> (2026).

Acknowledgements

We would like to thank A. Benjamin, K. Daruwalla, Y. Jo and I. Voito for feedback on the manuscript.

Author contributions

M.J., M.S., Y.A. and G.H. conceived the study. M.J., M.S. and G.H. performed the analyses, with feedback from Y.A. and S.B.H. M.J., M.S., G.H. and Y.A. wrote the manuscript, with feedback from S.B.H.

Funding

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (<https://www.csd3.cam.ac.uk>), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (<https://www.dirac.ac.uk>). M.S. was funded by an Engineering and Physical Sciences Research Council (EPSRC DTP) studentship (RG94782). S.B.H. and M.J. were supported by the Sainsbury Wellcome Centre core grant from the Gatsby Charitable Foundation and the Wellcome Trust (090843/F/09/Z) and a Wellcome Investigator Award (S.B.H., 219561/Z/19/Z). M.J. was additionally supported by the Cold Spring Harbor Laboratory Fellows Program and the Pershing Square Foundation. Y.A. was supported by UK Research and Innovation Biotechnology and Biological Sciences Research Council research grant BB/X013235/1. The funders had no role in study

design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

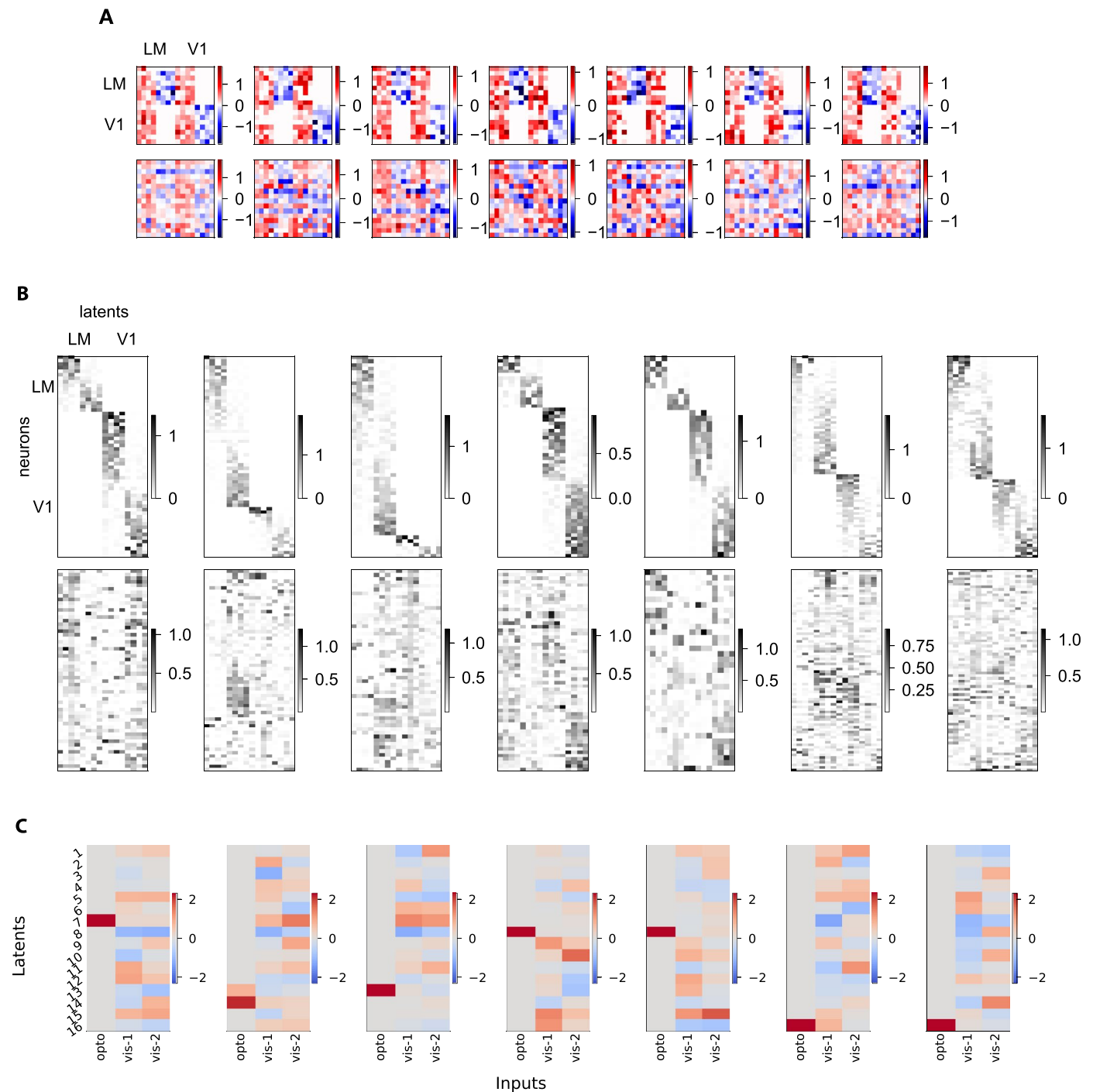
Extended data is available for this paper at <https://doi.org/10.1038/s41593-026-02437-3>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-026-02437-3>.

Correspondence and requests for materials should be addressed to Mitra Javadzadeh or Marine Schimel.

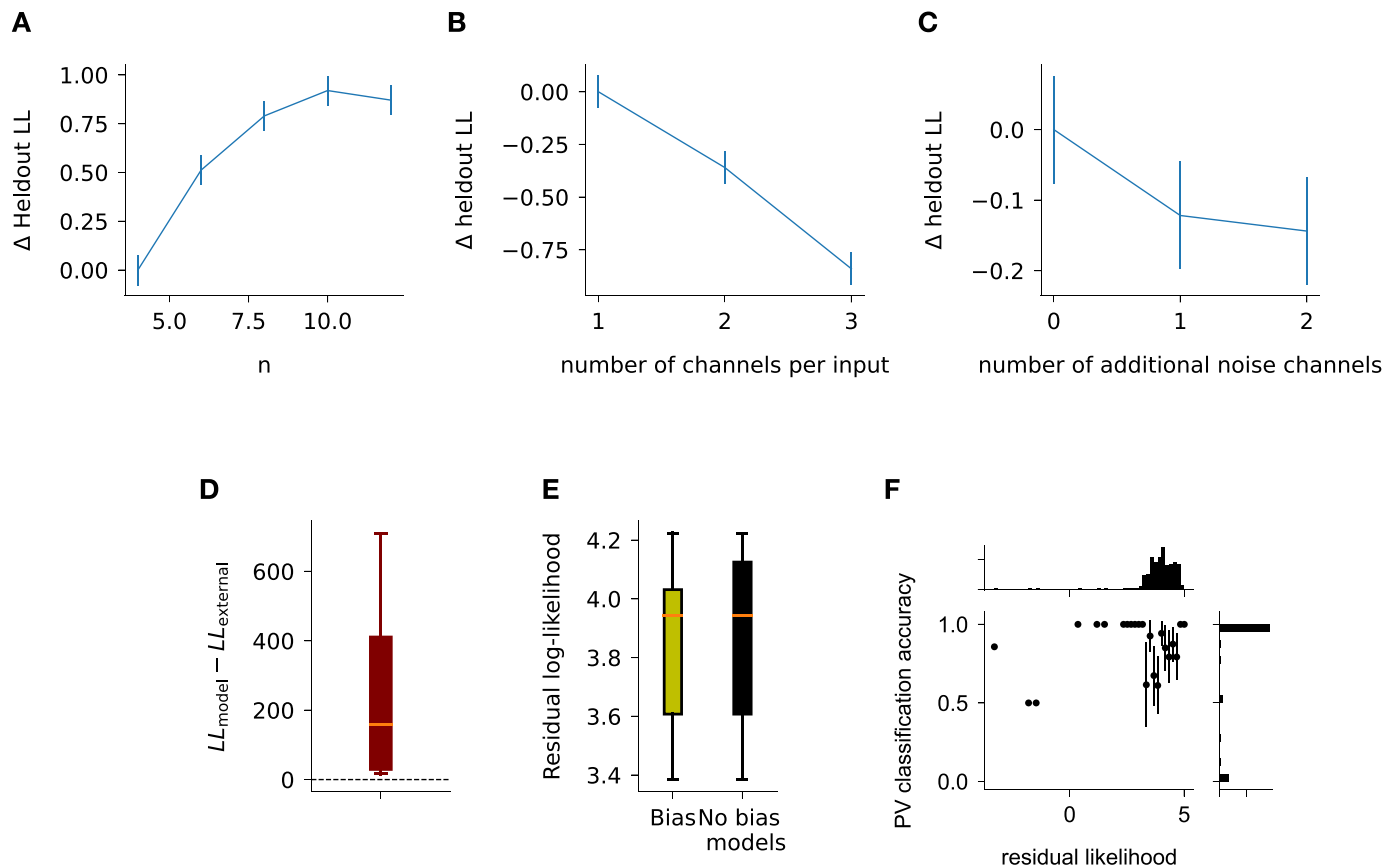
Peer review information *Nature Neuroscience* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.



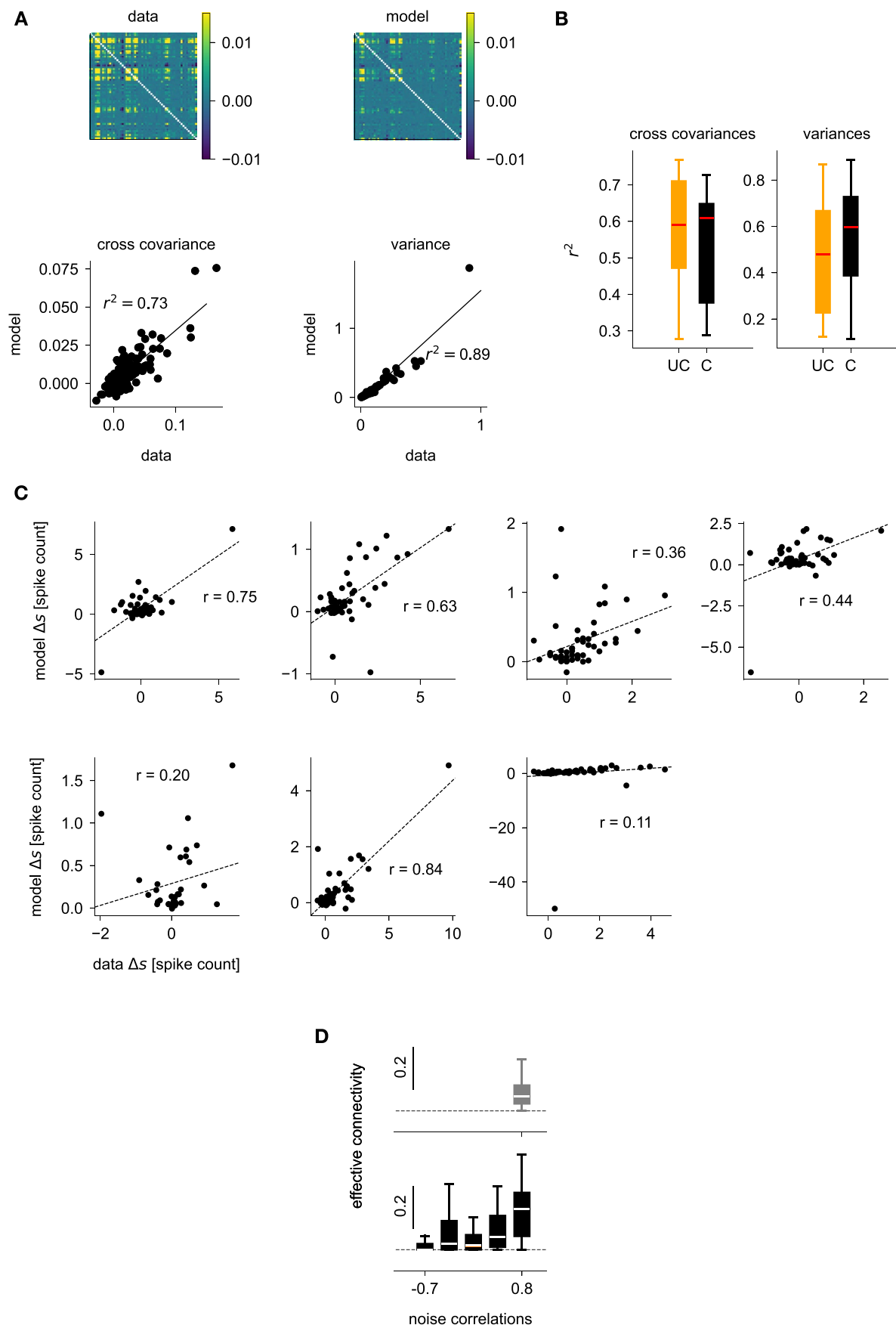
Extended Data Fig. 1 | W and C learned by the constrained and unconstrained models. (A) Top: latent circuit connectivity, W , of the constrained models for all animals. Bottom: as in top, but for unconstrained models of the same animals. (B) Top: the readout matrix, C , of constrained models, as in Fig. 1f, but shown for

all animals. Bottom: as in top but for unconstrained models of the same animals. (C) Visualization of the learned input coefficients (B matrices) for all animals. The columns correspond to the different input channels (opto and visual stimuli) and the rows to the $n = 16$ latent units.



Extended Data Fig. 2 | Model selection. (A) Normalized residual heldout log-likelihood of models of different size plotted against n , the number of latent units per area. Normalization was done by subtracting the average residual log-likelihood (across animals, cells, and cross validation folds) of the smallest model ($n = 4$). (B) Normalized residual heldout log-likelihood of models of size $n = 16$, as we increase the number of input channels. In this case, we used a stimulus-dependent prior, and used either $m = 1$, $m = 2$, or $m = 3$ input channels per stimulus (go, no-go, and opto). As in (A), normalization was done by subtracting the average residual log-likelihood (across animals, cells, and cross validation folds) of the smallest model ($m = 1$). (C) Normalized residual heldout log-likelihood of models of size $n = 16$, as we increase the number of input channels. Here, we used 3 channels with the standard stimulus-dependent prior, and we used a flat prior for each additional channel. As in (A-B), normalization was done by subtracting the average residual log-likelihood (across animals, cells, and cross validation folds) of the smallest model ($m_{\text{extra}} = 0$). In (A-C), each point represents the mean across neurons, animals, and train/test splits of the data ($n = 738$). Error bars denote $\pm 2 \times \text{s.e.m.}$ (D) Difference in the residual heldout log-likelihood of our

standard models, and deterministic models driven by condition-dependent inputs that we set to be equal to the envelope of our prior. While we trained the deterministic models using the same objective as our standard models, their log-likelihood is considerably worse, as they are unable to capture any trial-to-trial variability. Each point represents one animal ($n = 7$). Box plots show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. (E) Residual heldout log-likelihood of models trained with (left) and without (right) a bias term. For each setting, we selected the best model per animal. Each point represents one animal ($n = 7$). Bias versus no-bias comparisons were assessed using a paired two-sided t -test ($p = 0.01$). Box plots show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. (F) The PV+ classification accuracy plotted against residual likelihoods for all trained models across all animals (each point represents one model, 239 models overall). Error bars are $\pm 2 \times \text{s.e.m.}$ Marginal distributions are shown on top and right.

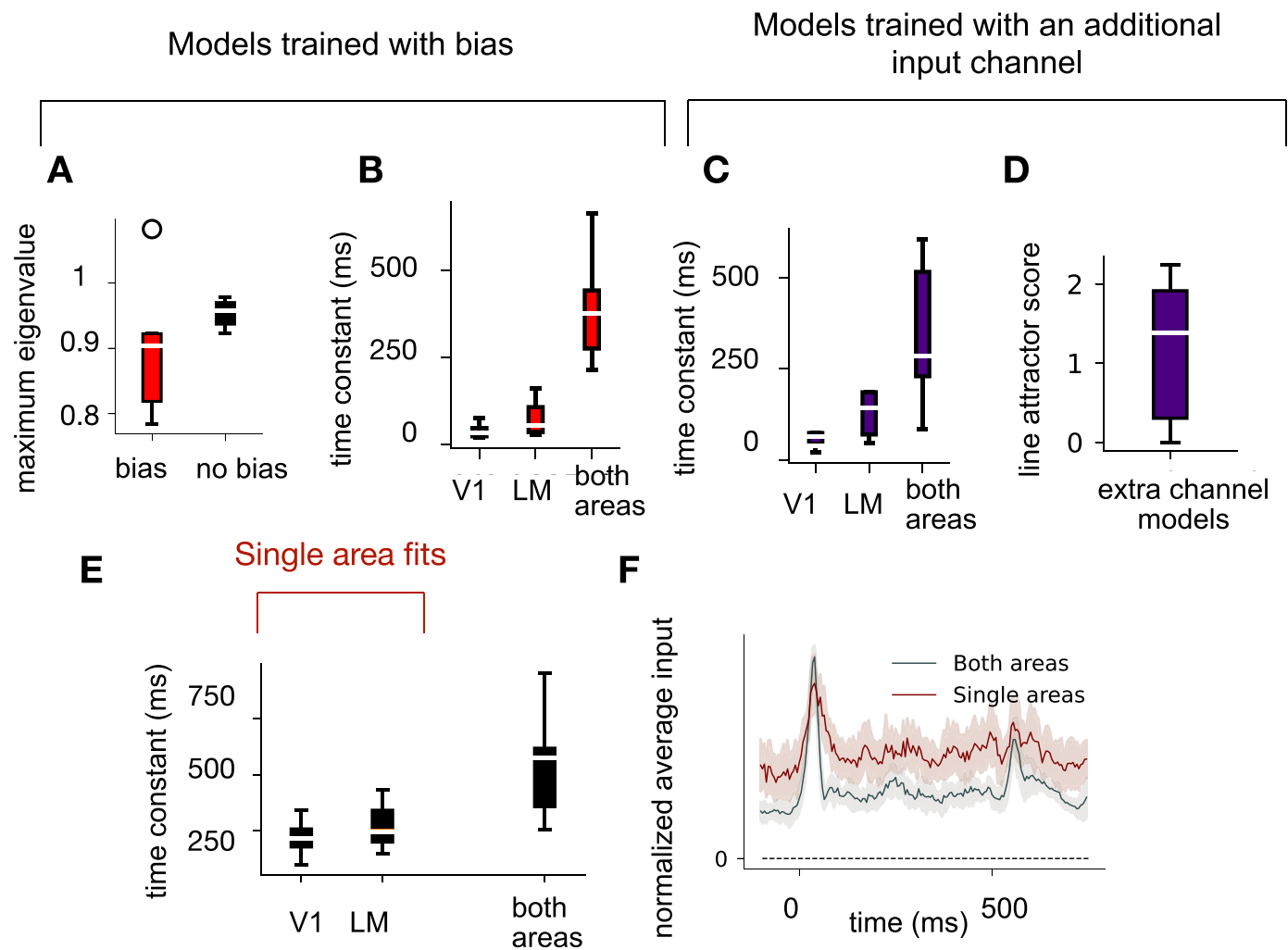


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Model performance on predicting covariances.

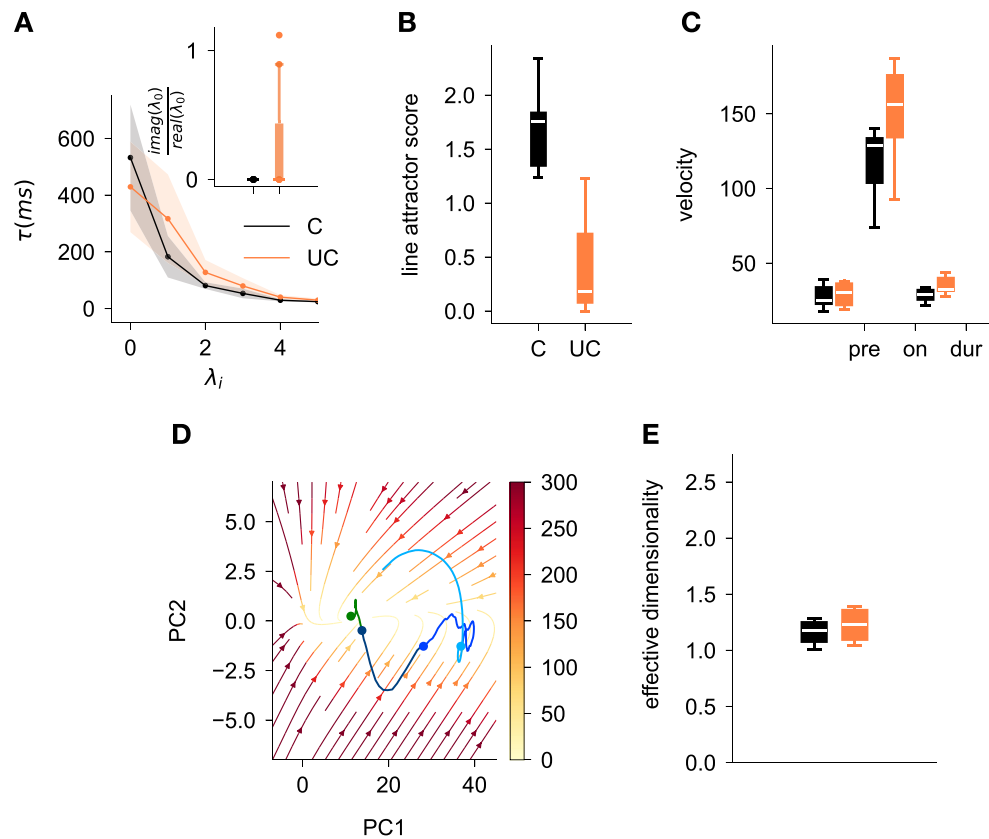
(A) Top: Covariance matrices of empirical neural firing (left) and model-predicted neural activity (right). Diagonal elements (variances) are not shown. Bottom left: Off-diagonal elements of the model predicted covariance matrix, plotted against the corresponding elements in the data covariance matrix, for one example animal. Each point represents a pair of neurons ($n=3540$ pairs). Bottom right: Empirical variance of neural activity plotted against the model-predicted ones, for the same example animal as in left. Each point represents one neuron ($n=60$ neurons). The model in (A) refers to the circuit-constrained model. **(B)** Distribution over animals of the coefficient of determination (r^2 , model vs. data) for the off-diagonal elements of the covariance matrix (left) and the variances (right). The constrained models are shown in black and the

unconstrained ones in orange. Each point represents one model ($n=7$ constrained and $n=6$ unconstrained). Constrained vs. unconstrained differences are not significant (two-sided Wilcoxon rank-sum test). Box plots show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. **(C)** As in Fig. 2d middle, but separately for each animal, using the same condition and optogenetic perturbation time window as the main figure. In order, $n = 56, 59, 53, 54, 36, 73, 74$ neurons. **(D)** As in Fig. 2g (main text), but using go instead of no-go trials. Each point represents a pair of latent units ($n=112$ pairs from 7 animals). Box plots show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range.



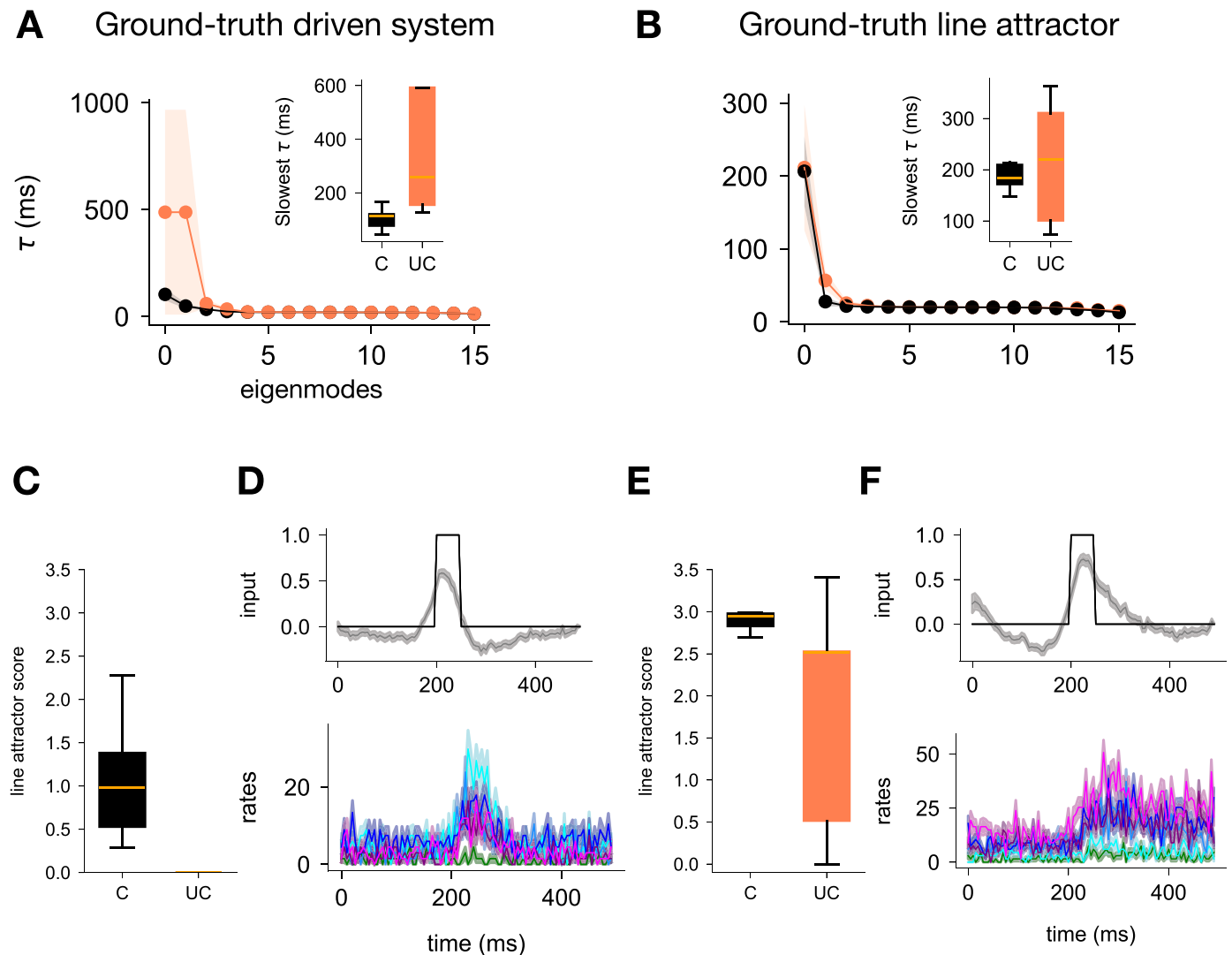
Extended Data Fig. 4 | Attractor controls. (A): Comparison of the maximum eigenvalue of models whose bias was learned from the start (red), relative to models whose bias term was fixed to be 0 (black). We trained multiple models per animal (15 bias models and 50 no-bias models) and selected the one with the highest residual log-likelihood (see details in Methods). Each point represents one model ($n=7$). Paired two-sided t -tests for bias vs. no bias comparison, $p=0.2$. (B): Distribution of the slowest timescale of individual areas, or the full dynamics, in models whose bias term was then allowed to be learned after 12000 iterations. For each animal, this analysis was performed for the model with the highest residual log-likelihood in the absence of bias (i.e. the models analysed in the main text). Each point represents one model ($n=7$). Comparisons between groups were assessed using paired two-sided t -tests. Full network vs V1: $p=0.0009$; full network vs LM: $p=0.001$; V1 vs LM: $p=0.08$. (C): Distribution of the slowest timescale of individual areas, or the full dynamics, in models with 1, 2, or 3 additional input channels. For each animal, we trained 3 models per number of additional input channels, and selected the models that had the highest residual log-likelihood across that channel number and identified PV+ cells as inhibitory. This led to a total of $n=9$ models across all animals. Each

point represents one model ($n=9$). Comparisons between groups were assessed using two-sided t -tests. Full network vs V1: $p=0.003$; full network vs LM: $p=0.43$; V1 vs LM: $p=0.25$. (D): Distribution of line attractor scores of models with 1, 2, or 3 additional input channels (selected in the same way as in C). Each point represents one model ($n=9$). (E): Slowest timescale of the linearized dynamics in models fit to data from individual areas (V1 or LM), or to the whole network (Both areas). For each animal, we fit models to the area that received optogenetic perturbation, and selected one model using the selection procedure described in Methods, yielding $n=4$ V1-only models, $n=3$ LM-only models, and $n=7$ full-network models. Comparisons between groups were assessed using two-sided independent t -tests. Full network vs V1: $p=0.03$; full network vs LM: $p=0.13$; V1 vs LM: $p=0.55$. (F): Average norm of the input to the latent circuit, for models fit to individual areas (selected in the same way as in E) in maroon, and models fit to the full circuit in gray. We normalize the norm to its maximal value across time for every trial, and take the average across trials for every animal. Lines denote the mean across models and shaded areas represent $\pm 2 \times \text{s.e.m.}$ across models ($n=7$). Box plots in A–E show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range.



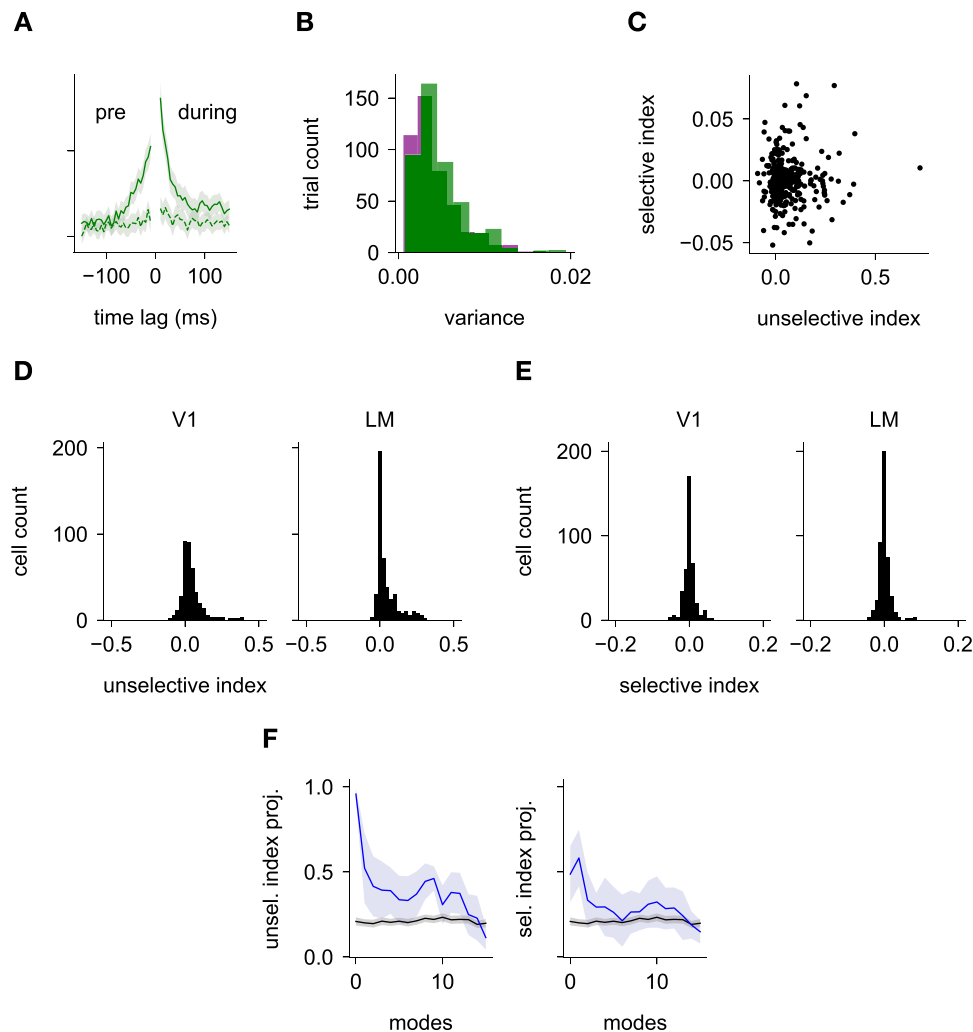
Extended Data Fig. 5 | Unconstrained models also recover almost-one-dimensional slow dynamics. (A–C) As in Fig. 3f, 3g, and 3e, respectively, but additionally showing unconstrained models in orange. One model per animal was selected for analysis ($n=7$ constrained and $n=6$ unconstrained models). (D) Flow field of the dynamics and the inferred state-space trajectories averaged over no-go trials, as in Fig. 3d, but for the only unconstrained model with a 0 line attractor score. This model exhibited rotations, but on very slow time constants relative to the 0.5s stimulus duration ($2\pi\tau_0/|\text{Im}(\lambda_0)| \approx 1.7$ s, where λ_0 is the eigenvalue with the largest real part, and $\tau_0 = 20$ ms is the single neuron time

constant). (E) Effective dimensionality of latent trajectories during the no-go stimulus (100–500 ms) projected on the plane spanned by the first two eigenvectors of the jacobian of the dynamics, which correspond to the two slowest eigenvalues. Effective dimensionality was calculated as the participation ratio of the covariance matrix: $d = \frac{1}{\sum_i p_i^2}$, where $p_i = \frac{\lambda_i}{\sum_j \lambda_j}$, and λ_i are the eigenvalues of the covariance matrix. One model per animal was selected for analysis ($n=7$ constrained and $n=6$ unconstrained models).



Extended Data Fig. 6 | Models fit to synthetic data recover ground-truth dynamical regimes. (A–B) Timescales of unconstrained (orange) and constrained (black) two-area models fitted to synthetic datasets in which the ground truth was either (A) a system without long-range connections and therefore no slow dynamics, but driven by a slow external input with 400 ms autocorrelation time shared across the two areas, or (B) a “non-driven” system where long-range connections establish an approximate line attractor with a single slow timescale of 400 ms. In both cases, a stimulus was delivered consistently in each trial at $t = 200$ ms, in the form of a transient 50-ms input pulse. We also artificially silenced either V1 or LM for 150 ms at random times in one third of the trials. Each point in box

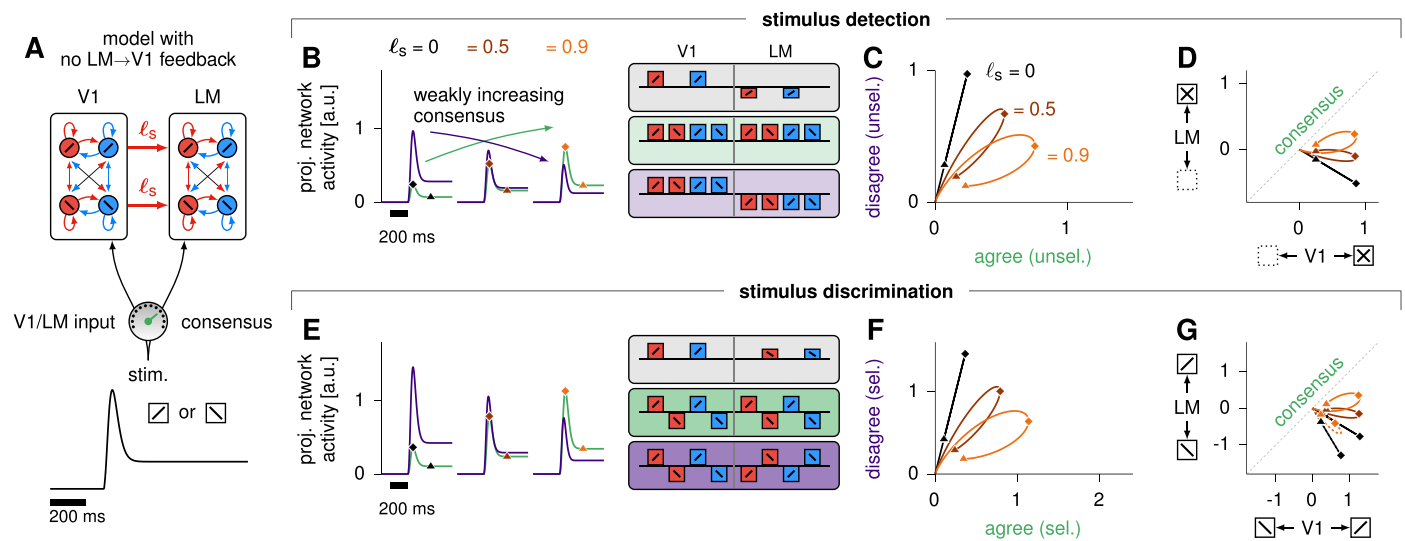
plots represents one model ($n = 7$ models, one per animal). Shaded areas represent $\pm 2 \times$ s.e.m. across models. (C) Distribution across animals of the attractor score of the recovered models whose timescales are shown in (A). Each point represents one animal ($n = 7$) (D) Top: In gray, average visual input for one example animal for the model shown in (A), overlaid with the true visual stimulus (black). Bottom: For the same animal, trial-averaged activity of example 4 neurons. (E–F) Same as (C–D) for the model shown in (B). See Section S2 for modeling details. Box plots in A–C and E show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range. Shaded areas represent $\pm 2 \times$ s.e.m. across trials.



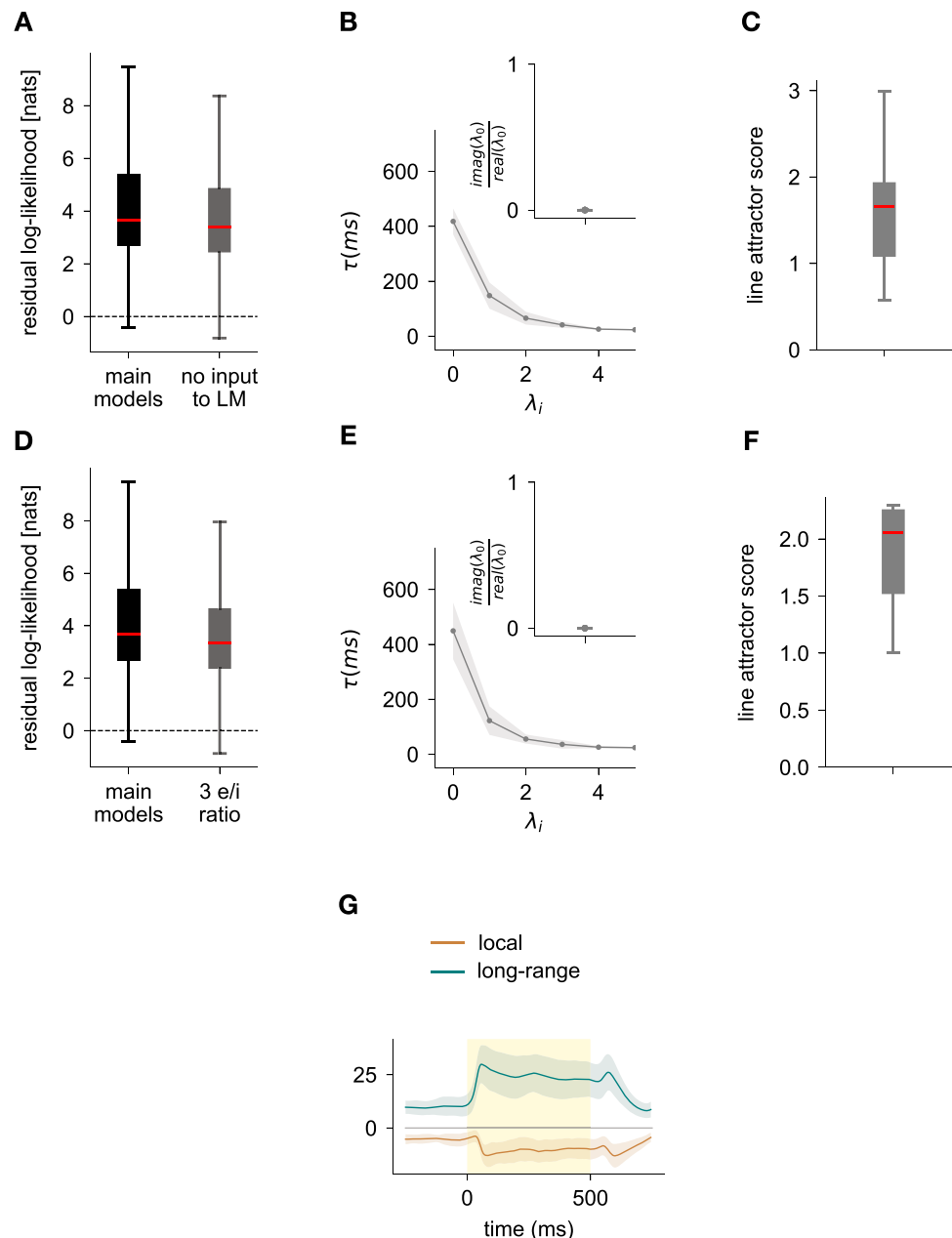
Extended Data Fig. 7 | Controls for selective and unselective modes.

(A) Autocorrelation function of model predicted neural activity pre- (left half) and during stimulus (right half), projected onto the 'agree-balanced' (solid) and 'agree-unbalanced' (dashed) modes. **(B)** Distribution over trials of the total variance of neural activity during the stimulus, projected onto the 'agree-balanced' (green) and 'disagree-balanced' (purple) modes. Median difference (agree-disagree) = 0.0005, $p < 10^{-5}$ from a two-sided Wilcoxon signed-rank test. **(C)** Distribution of the selective indices across V1 and LM in all animals, against the nonselective indices. Pearson $r = 0.05$, $p > 0.05$. **(D)** Distribution of the unselective indices across all neurons and animals, separately for V1 (left) and LM (right) neurons (V1 mean = 0.046, $p < 10^{-5}$, LM mean = 0.047, $p < 10^{-5}$). P-values are two-sided t-tests. **(E)** Same as (D) but for selective indices

(V1 mean = 2×10^{-5} , $p > 0.05$, LM mean = 0.0008, $p > 0.05$). **(F)** Alignment of the latent circuits' eigenmodes onto agree unselective (left) and selective modes (right) in blue (see Methods). The eigenmodes on the x-axis are ordered from the slowest (longest decay time constant) to fastest, and the y-axis shows the absolute value of the dot products between the eigenmodes and the unit-length vectors containing latent circuit's unselective or selective indices. Shaded areas represent $\pm 2 \times$ s.e.m. The black lines denote chance level ($\pm 95\%$ confidence intervals), resulting from projecting random unit-length (16D) vectors onto the eigenmodes. The elements of the random vectors were drawn from a Gaussian distribution ($\mu = 0$, $\sigma = 1$), and 20 random vectors per animal were used to calculate confidence intervals. All analyses were performed on the subset of animals for which latent circuit models were fit ($n = 7$ animals).



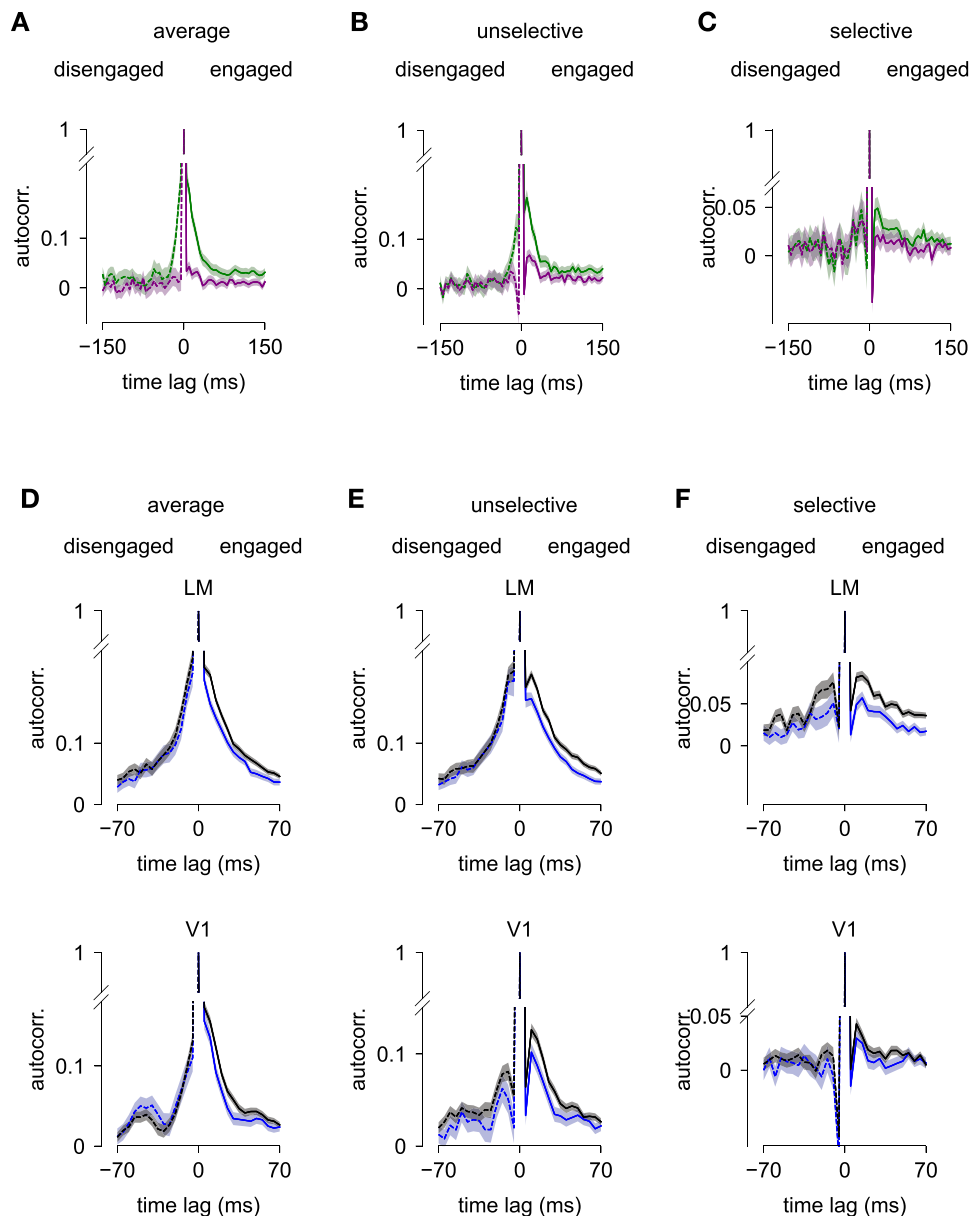
Extended Data Fig. 8 | Minimal model without feedback connections from LM to V1. (A–G) As in Fig. 6a–g (main text), but with no feedback connections from LM to V1.



Extended Data Fig. 9 | Robustness of results to various model design choices.

(A–C) Models with external inputs targeting only V1 (see Methods). One model per animal was selected for analysis ($n=7$). (A) Residual log-likelihood of the predicted firing rates, as in Fig. 2b, for the main models (black), compared to models no external input to LM (gray). Two-sided Wilcoxon signed-rank test, $p < 0.001$. (B) As in Fig. 3f, but for models with no external input to LM. (C) As in Fig. 3g, but for models no external input to LM. (D–F) Same plots as (A–C), but for

models with a 3/1 ratio of excitatory to inhibitory latent units (see Methods). One model per animal was selected for analysis ($n=7$). (D) Two-sided Wilcoxon signed-rank test, $p < 0.001$. (G) Local (bronze) and long-range (teal) inputs to the V1-LM circuit, as in Fig. 3b left, but for models with a 3/1 ratio of excitatory to inhibitory latent units. Shaded areas represent $\pm 2 \times \text{s.e.m.}$ Box plots in A–F show the median (center line), interquartile range (box bounds), and whiskers extending to the farthest points within $1.5 \times$ the interquartile range.



Extended Data Fig. 10 | Autocorrelation analyses in engaged and disengaged trials. (A) Right half reproduced from Fig. 4h (right half). Left half, as in right half but for disengaged trials (see Methods). (B) As in (A) but for neural activity projected onto the agree/disagree unselective modes (see Methods). (C) As in (A) but for neural activity projected onto the agree/disagree selective modes (see

Methods). See Section S3 and Table S2 for the estimated timescales across all engaged and disengaged conditions. (D-F) All right halves reproduced from Fig. 5a-c. Left halves as in right, but for disengaged trials (see Methods). See Section S3 and Table S1 for the estimated timescales across all engaged and disengaged conditions.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Custom code used for data collection is described in a previously published manuscript (Javadzadeh and Hofer, Neuron 2022).

Data analysis Code used to fit the iLQR-VAE models is available at https://github.com/marineschimmel/v1lm_nn.
The repository includes model-fitting code and instructions for running the model-fitting pipeline as well as some example analyses.
Fitted models used in the manuscript are available on Figshare at <https://doi.org/10.6084/m9.figshare.31142140>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Fitted iLQR-VAE models used for the model-based analyses, including the training and test datasets, are available on Figshare at <https://doi.org/10.6084/>

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Data exclusions

Replication

Randomization

Blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a ☐ Involved in the study

☒ ☐ Antibodies

☒ ☐ Eukaryotic cell lines

☒ ☐ Palaeontology and archaeology

☐ ☒ Animals and other organisms

☒ ☐ Clinical data

☒ ☐ Dual use research of concern

☒ ☐ Plants

Methods

n/a ☐ Involved in the study

☒ ☐ ChIP-seq

☒ ☐ Flow cytometry

☒ ☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals	Details of the experimental data used for this study are included in a previously published manuscript (Javadzadeh and Hofer, Neuron 2022).
Wild animals	No wild animals were used.
Reporting on sex	Both male and female were used in the study.
Field-collected samples	No field collected samples were used in the study.
Ethics oversight	Details of the experimental data used for this study are included in a previously published manuscript (Javadzadeh and Hofer, Neuron 2022).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>