State-dependent control of cortical processing speed via gain modulation

David Wyrick[a] **and Luca Mazzucato**[a,b]

[a] *Department of Biology and Institute of Neuroscience,* [b] *Department of Mathematics and Physics, University of Oregon, Eugene.*

E-mail: lmazzuca@uoregon.edu

Abstract: To thrive in dynamic environments, animals can generate flexible behavior and rapidly adapt responses to a changing context and internal state. Examples of behavioral flexibility include faster stimulus responses when attentive and slower responses when distracted. Contextual modulations may occur early in the cortical hierarchy and may be implemented via afferent projections from top-down pathways or neuromodulation onto sensory cortex. However, the computational mechanisms mediating the effects of such projections are not known. Here, we investigate the effects of afferent projections on the information processing speed of cortical circuits. Using a biologically plausible model based on recurrent networks of excitatory and inhibitory neurons arranged in cluster, we classify the effects of cell-type specific perturbations on the circuit's stimulus-processing capability. We found that perturbations differentially controlled processing speed, leading to counter-intuitive effects such as improved performance with increased input variance. Our theory explains the effects of all perturbations in terms of gain modulation, which controls the timescale of the circuit dynamics. We tested our model using large-scale electrophysiological recordings from the visual hierarchy in freely running mice, where a decrease in single-cell gain during locomotion explained the observed acceleration of visual processing speed. Our results establish a novel theory of cell-type specific perturbations linking connectivity, dynamics, and information processing via gain modulations.

Contents

1 Introduction

Animals respond to the same stimulus with different reaction times depending on the context or the behavioral state. Faster responses may be elicited by expected stimuli or when the animal is aroused and attentive [1]. Slower responses may occur in the presence of distractors or when the animal is disengaged from the task [2–4]. Experimental evidence suggests that neural correlates of these contextual modulations occur early in the cortical hierarchy, already at the level of the primary sensory cortex [5, 6]. During the waking

state, levels of arousal, attention, and task engagement vary continuously and are associated with ongoing and large changes in the activity of neuromodulatory systems [7–9] as well as cortico-cortical feedback pathways [10–14]. Activation of these pathways modulate the patterns of activity generated by cortical circuits and may affect their information-processing capabilities. However, the precise computational mechanism underlying these flexible reorganizations of cortical dynamics remains elusive.

Variations in behavioral and brain state, such as arousal, engagement and body movements may act on a variety of timescales, both slow (minutes, hours) and rapid (seconds or subsecond), and spatial scales, both global (pupil diameter, orofacial movements) and brain subregion-specific; and they can be recapitulated by artificial perturbations. These variations have been associated with a large variety of seemingly unrelated mechanisms operating both at the single cell and at the population level. At the population level, these mechanisms include modulations of low and high frequency rhythmic cortical activities [15]; changes in noise correlations [16, 17]; and increased information flow between cortical and subcortical networks [15]. On a cellular level, these variations have been associated with modulations of single-cell responsiveness and reliability [17]; and cell-type specific gain modulation [15]. These rapid, trial-by-trial modulations of neural activity may be mediated by neuromodulatory pathways, such as cholinergic and noradrenergic systems [7–9, 18], or more precise cortico-cortical projections from prefrontal areas towards primary sensory areas [10–14]. The effects of these cortico-cortical projections can be recapitulated by optogenetic activation of glutamatergic feedback pathways [19]. In the face of this wide variety of physiological pathways, is there a common computational principle underlying the effects they elicit on sensory cortical circuits?

A natural way to model the effect of activating a specific pathway on a downstream circuit is in the form of a perturbation to the downstream circuit's afferent inputs or recurrent couplings [20, 21]. Here, we will present a theory explaining how these perturbations control the information-processing speed of a downstream cortical circuit. Our theory shows that the effects of perturbations that change the statistics of the afferents or the recurrent couplings can all be captured by a single mechanism of action: intrinsic gain modulation. Our theory is based on a biologically plausible model of cortical circuits using clustered spiking network [22]. This class of models capture complex physiological properties of cortical dynamics such as state-dependent changes in neural activity, variability [23–27] and information-processing speed [20]. Our theory predicts that gain modulation controls the temporal dynamics of the cortical circuit and thus its information processing speed, such that decreasing the intrinsic single-cell gain leads to faster stimulus coding.

We tested our theory by examining the effect of locomotion on visual processing in the visual hierarchy. We found that locomotion decreased the intrinsic gain of neurons in the absence of stimuli in freely running mice. The theory thus predicted a faster encoding of visual stimuli during running compared to rest, which we confirmed to be the most prominent effect of locomotion on visual processing. Our theoretical framework links gain modulation to information-processing speed, providing guidance for the design and interpretation of future manipulation experiments by unifying the changes in brain state due to behavior, optogenetic, or pharmacological perturbations, under the same shared mechanism.

2 Results

2.1 Controlling information processing speed with perturbations

To elucidate the effect of perturbations on cortical networks, we modeled the local circuit as a network of recurrently connected excitatory (E) and inhibitory (I) spiking neurons. Both E and I populations were arranged in clusters[20, 22, 23, 25, 28], where synaptic couplings between neurons in the same cluster were potentiated compared to neurons in different clusters, reflecting the empirical observation of cortical assemblies of functionally correlated neurons (Fig. 1a-b, [29–32]). In the absence of external stimulation (ongoing activity), clustered networks generate rich temporal dynamics characterized by metastable activity operating in the inhibition stabilized regime (Fig. S1), with lognormal distributions of firing rates (Fig. 1c). Network activity was characterized by the emergence of the slow timescale of cluster transient activation with average activation lifetime of 106±35 ms (hereby referred to as “cluster timescale"), much larger than single neuron time constant (20ms) [23, 25].

To investigate the effect of afferent perturbations, we compared the network's information-processing speed by presenting stimuli in an unperturbed and a perturbed condition. In unperturbed trials (Fig. 2a), we presented one of four sensory stimuli, modeled as depolarizing currents targeting a subset of stimulus-selective E neurons. Stimulus selectivities were mixed and random, clusters having equal probability of being stimulus-selective. In perturbed trials, in addition to the same sensory stimuli, we included a perturbation, which was turned on before the stimulus and was active until the end of stimulus presentation. We investigated and classified the effect of several perturbations. The first type of perturbations

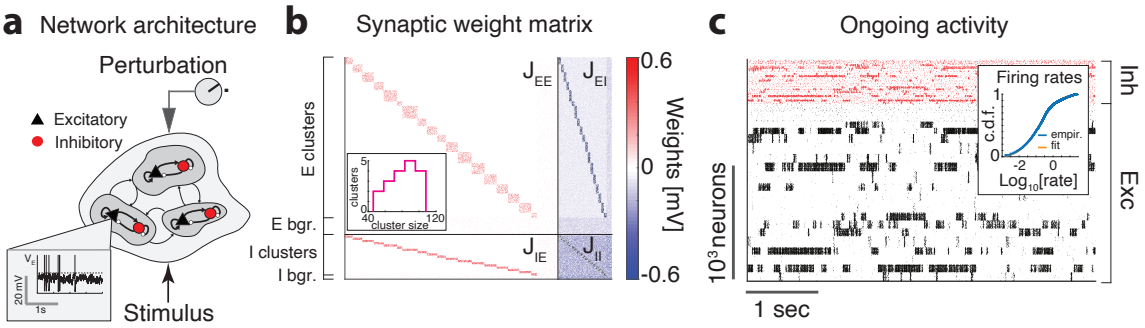

Figure 1. Biological plausible model of cortical circuit. **a**) Schematics of the network architecture. A recurrent network of E (black triangles) and I (red circles) spiking neurons arranged in clusters (inset: membrane potential trace from representative E neuron) is presented sensory stimuli targeting subsets of E clusters, in different conditions defined by perturbations. **b**) Synaptic couplings J_{ij} for a representative clustered network, highlighting the block diagonal structure of potentiated intra-cluster synaptic weights for both E and I clusters, and the background E and I populations (bgr). Cluster size was heterogeneous (inset). **c**) Representative ongoing trial; tick marks represent spike times of E (black) or I (red) neurons. Inset: The cumulative distributions of single-cell firing rates (**c**) in the representative network are lognormal (blue: empirical data; orange: lognormal fit).

δmean(E) or δmean(I) affected the mean of the afferent currents to either E (Fig. 2b) or I populations, respectively. In the second type of perturbation δvar(E) or δvar(I), we chose a cell-type specific population (either E or I, respectively), then for each neuron in that population we sampled its external current from a normal distribution with zero mean and fixed variance. This perturbation thus introduced a *spatial variance* in the cell-type specific afferent currents. In the third type of perturbations δAMPA or δGABA, we changed the average GABAergic or glutamatergic (AMPA) recurrent synaptic weights. Perturbations were identical in all trials of the perturbed condition; namely, they did not convey any information about the stimuli. We chose the range of external perturbations such that the network still retained non-trivial metastable dynamics within the whole range. Namely, we avoided extreme perturbations as unphysiological, where the network activity completely shut down or saturated loosing metastability.

We assessed how much information about the stimuli was encoded in the population spike trains at each moment using a multiclass classifier (with four class labels corresponding to the four presented stimuli, Fig. 2c). In the unperturbed condition, the time course of the cross-validated decoding accuracy, averaged across stimuli, was significantly above chance after $0.21 +/- 0.02$ seconds (mean$\pm$s.e.m. across 10 simulated networks, black curve in Fig. 2c) and reached perfect accuracy after a second. In the perturbed condition, stimulus identity was decoded at chance level in the period after the onset of the perturbation but before stimulus presentation (Fig. 2c), consistent with the fact that the perturbation did not convey information about the stimuli. We found that perturbations significantly modulated the network information processing speed. We quantified this modulation as the average latency to reach a decoding accuracy between 40% and 80% (Fig. 2c, yellow area), and found that perturbations differentially affected processing speed.

Perturbations had opposite effects depending on which cell-type specific populations they targeted. Increasing δmean(E) monotonically improved network performance (Fig. 2d, left panel): in particular, positive perturbations induced an anticipation of stimulus-coding (shorter latency), while negative ones led to longer latency and slower coding. The opposite effect was achieved when increasing δmean(I), which slowed down processing speed (Fig. 2d, right panel). Perturbations that changed the spatial variance of the afferent currents had counterintuitive effects (Fig. 2e). We measured the strength of these perturbations via their coefficient of variability $CV(\alpha) = \sigma_\alpha/\mu_\alpha$, for $\alpha = E, I$, where σ and μ are the standard deviation and mean of the across-neuron distribution of afferent currents. Perturbations δvar(E) that increased $CV(E)$ led to faster processing speed. The opposite effect was achieved with perturbations δvar(I) inducing a spatial variance across afferents to I neurons, which slowed down stimulus-processing speed (Fig. 2g). Perturbations δAMPA which increased the glutamatergic synaptic weights improved performance proportionally to the perturbation. The opposite effect was achieved by perturbations δGABA that increased the GABAergic synaptic weights, which monotonically decreased network processing speed (Fig. 2g). We thus concluded that afferent current perturbations differentially modulated the speed at which network activity encoded information about incoming sensory inputs. Such modulations exhibited a rich dynamical repertoire.

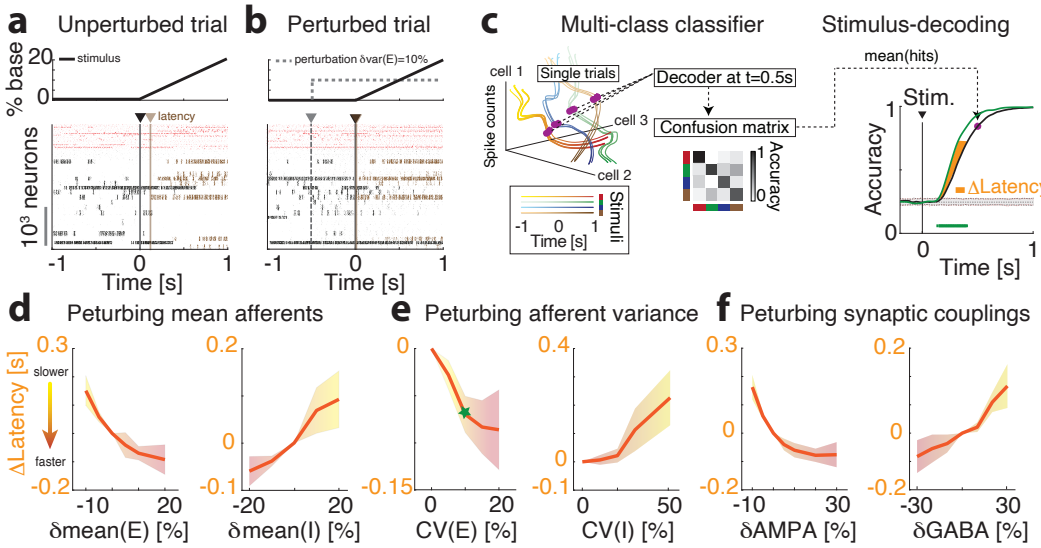

Figure 2. Perturbations control stimulus-processing speed in the clustered network. **a-b**) Representative trials in the unperturbed (**a**) and perturbed (**b**) conditions; the representative perturbation is an increase in the spatial variance δvar(E) across E neurons (corresponding to the green decoding curve in **c** and the green star in **e**). After a stimulus is presented at $t = 0$ (black vertical line overlaid on raster plot represent stimulus onset; the black curve in the top panel represent the stimulus time course), stimulus-selective E-clusters (brown tick marks represent their spiking activity) are activated at a certain latency (brown vertical line). In the perturbed condition (**b**), a perturbation is turned on before stimulus onset (gray-dashed vertical line overlaid on raster plot represent perturbation onset; the same line in the top panel represent the perturbation time course). The activation latency of stimulus-selective clusters is shorter in the perturbed compared to the unperturbed condition. **c**) Left: schematic of stimulus-decoding analysis. Left: A multi-class classifier was trained to discriminate between the four stimuli from single-trial population activity vectors in a given time bin (left: (curves represent the time course of population activity in single trials, color-coded for 4 stimuli; the purple circle highlights a given time bin along the trajectories), yielding a cross-validated confusion matrix for the decoding accuracy at that bin (central panel). Right: Average stimulus-decoding accuracy in each bin in the unperturbed (black curve) and perturbed (green curve) conditions (horizontal green bar: significant different between conditions, $p < 0.05$ with multiple bin correction). **d-g**: Difference in stimulus decoding latency in the perturbed minus the unperturbed conditions (average difference between rise time of decoding accuracy in **c**, mean±S.D. across 10 networks) for six perturbations (see Methods and main text for details; green star represent the perturbation in **b-c**).

2.2 Single-cell responses cannot explain the effects of perturbations

What is the neural mechanism underlying the observed modulations of processing speed induced by perturbations? We investigated whether the effect of perturbations on processing speed could be captured by changes in single-cell responses. We first characterized single-cell responses to perturbations alone, in the absence of sensory stimuli (Fig. 3a). We found that perturbations differentially affected neuronal responses in a cell-type specific way. Perturbations changed the average population firing rates, and led to complex patterns of response

across E and I populations (Fig. 3b). Specifically, perturbations increasing δmean(E) induced higher firing rates and induced proportionally excited (inhibited) responses in both E and I populations. On the other hand, perturbations that increased δmean(I) led to a decrease in both E and I average firing rates (Fig. 3b). This paradoxical effect [33] revealed that the network operates in the inhibition stabilized regime. When increasing the inhibitory current beyond δmean(I)=50%, the network reached a reversal point where the E population activity became silent and the I population rebounded, starting to increase their firing rates again (Fig. S1).

Perturbations increasing the variance δvar(E) and δvar(I) led to surprising effects (Fig. 3b, S2a). Increasing δvar(E) induced higher firing rates in both E and I populations, despite leaving the mean afferents unchanged; moreover, it led to mixed responses at the single cell level, with a prevalence of excited responses in both E and I populations. We will see below that this set of responses is consistent with locomotion-induced effects in the visual cortical hierarchy. Increasing δvar(I) left I firing rates unchanged but led to a decrease of E firing rates. This perturbation also induced mixed responses at the single cell level, with a prevalence of inhibited responses in both populations. Finally, perturbations δAMPA and δGABA led to responses similar to those found when driving the mean E- or I-afferents, respectively.

We then hypothesized that, if the response increase were larger for stimulus-selective compared to nonselective neurons (i.e., if ΔPSTH(sel)>ΔPSTH(nonsel)), then a perturbation increasing single-cell stimulus-responses could lead to faster stimulus-processing speed. We found that perturbations strongly affected the peak of single-cell responses to stimuli compared to baseline; in particular, they affected stimulus-selective and non-selective neurons differentially (Fig. 3b, S2). We then examined changes in single-cell selectivity, measured by the d' of their responses across stimuli. We found that perturbations strongly affected single-cell selectivities with significant changes in their d' (Fig. 3c, S3).

We then tested whether perturbation-induced changes in stimulus responses or discriminability could consistently explain the entirety of the observed changes in stimulus-processing speed, namely be the mechanism mediating that effect. Changes in selectivity and responsiveness were consistent with modulation of processing speed in the case of perturbations targeting I populations, namely δmean(I), δvar(I), and δGABA, but they were inconsistent with modulations in processing speed when perturbation targeted E populations (Fig. S2c-d, S2b-c). In the case of the perturbation δvar(E), network performance increased with larger perturbations even though single-cell responses and selectivity increasingly degraded. In the case of the perturbation δmean(E) and δAMPA, network performance likewise increased but single-cell metrics where non-monotonic in the value of the perturbation. Across all different perturbations, changes in single-cell properties were overall inconsistent and accounted for a small fraction of the variance in modulations of processing speed (Fig. 3c-d). In conclusion, since changes in single-cell stimulus-responses across all perturbations were overall inconsistent with the observed changes in processing speed, we conclude that they could not represent the mechanism underlying the observed effects of perturbations.

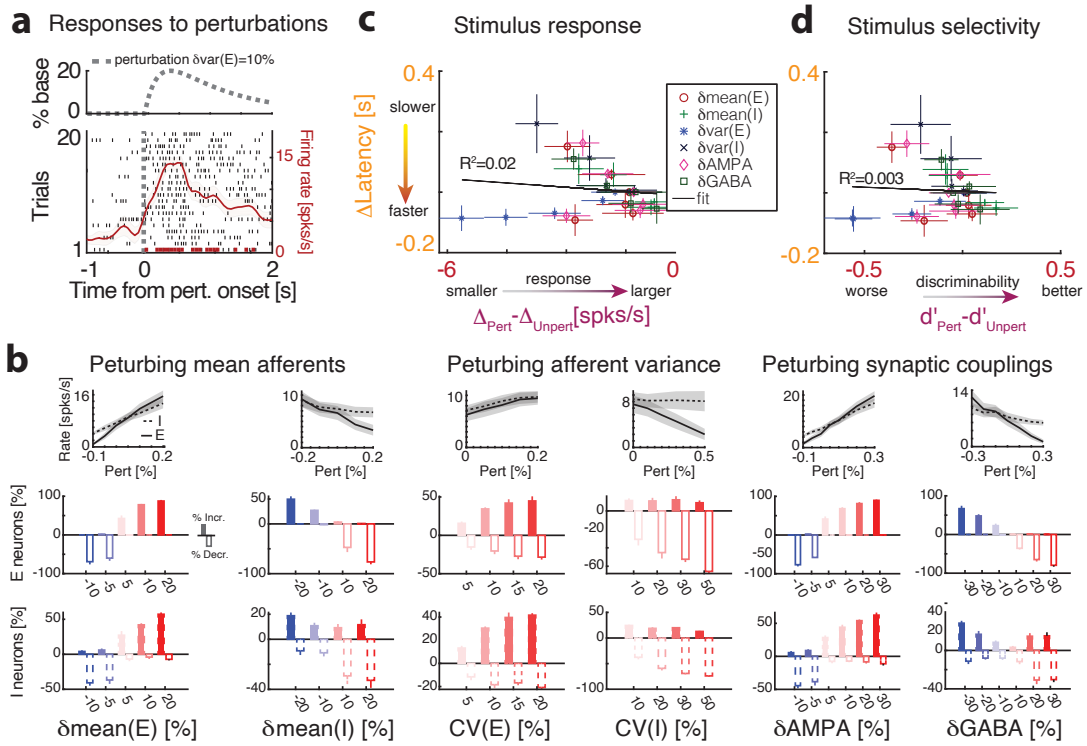

Figure 3. Single-cell responses to stimuli and perturbations. **a**) Representative single cell response to the perturbation δvar(E)=10% in the absence of stimuli (top: dashed line, time course of perturbation; bottom: dashed line, perturbation onset; red curve, response PSTH, mean±S.D. across 20 trials; horizontal red bar, significant response, $p < 0.05$ with multiple bin correction. **b**) Firing rate responses induced by the perturbations. Top: Average change in firing rate across E (full) and I (dashed) populations (mean±S.D. across 10 simulated networks). Histograms: Average fractions of E (top) and I (bottom) neurons with excited (positive bars) and inhibited (negative bars) responses to the perturbations (t-test, $p < 0.05$). **c**) Single-cell changes in firing rate response to stimuli due to the perturbations (Δ =peak response-baseline in each perturbed or unperturbed condition) are uncorrelated to changes in stimulus-decoding latencies (same as in Fig. 2d-f). Color-coded markers represent different perturbations, each point is the mean±s.e.m. across 10 networks for a specific value of the perturbation; linear regression, $R^2 = 0.02$. **d**) Single-cell changes in stimulus selectivity due to the perturbations (d') are uncorrelated to changes in stimulus-decoding latencies (same notation as in **c**; linear regression, $R^2 = 0.003$).

2.3 Modulations of the cluster timescale explain changes in stimulus-processing speed

A crucial feature of neural activity in clustered networks is metastable attractor dynamics, characterized by the emergence of a long timescale of cluster activation (Fig. 1c). We reasoned that, if the perturbations affected the intrinsic timescale of metastable dynamics, this could lead to changes in stimulus-processing speed. We first tested whether perturbations modulated the network's attractor dynamics. To isolate the effects of perturbations on the

attractor landscape, we considered a stimulation protocol where perturbations occurred in the absence of sensory stimuli ("ongoing activity"). We found that perturbations strongly modulated the attractor landscape, changing the repertoire of attractors the network activity visited during its itinerant dynamics (Fig. 4a-b).

Changes in attractor landscape were perturbation-specific. Perturbations increasing δmean(E) (δmean(I)) induced a consistent shift in the repertoire of attractors: larger perturbations led to larger (smaller) numbers of co-active clusters. Surprisingly, perturbations that increased δvar(E) (δvar(I)), led to network configurations with larger (smaller) sets of co-activated clusters. This effect occurred despite the fact that such perturbations did not change the mean afferent input to the network. Perturbations affecting δAMPA and δGABA had similar effects to δmean(E) and δmean(I), respectively.

We then tested whether perturbations modulated the network's intrinsic timescale of cluster activation. Indeed, we found that perturbations differentially modulated the average cluster activation timescale τ during ongoing periods, in the absence of stimuli (Fig. 4c). In particular, increasing δmean(E), δvar(E), or δAMPA led to a proportional acceleration of the network metastable activity and shorter τ; while increasing δmean(I), δvar(I) or δGABA induced the opposite effect with longer τ. Changes in τ were congruent with changes in the duration of intervals between consecutive activations of the same cluster (cluster inter-activation intervals, Fig. S3).

In all conditions, the perturbation-induced changes of the cluster timescale τ, estimated during ongoing periods, predicted the effect of the perturbation on stimulus-processing latency (Fig. 4c-d). Specifically, perturbations that induced an acceleration of τ in turn accelerated stimulus coding, and vice versa. This led us to formulate the following hypothesis for the computational mechanism underlying the modulation of the stimulus-processing speed. After stimulus presentation, network activity encodes stimulus-related information by activating the stimulus-selective clusters. If perturbations alter the onset latency of stimulus-selective clusters, they would control the latency at which the network activity starts encoding the stimulus. We can visualize this hypothesis in representative trials where the same stimulus was presented in the absence (Fig. 2a) or in the presence (Fig. 2b) of the perturbation δmean(E)= 10%. Stimulus-selective clusters (highlighted in brown) had a faster activation latency in response to the stimulus in the perturbed condition compared to the unperturbed one. A systematic analysis confirmed this hypothesis revealing that, for all perturbations, the activation latency of stimulus-selective clusters was the best predictor of the change in decoding latency (Fig. 4e, $R^2 = 0.93$).

These results demonstrate that the effect of perturbations on stimulus-processing speed originates in their modulations of the metastable dynamics of cluster activation. In particular, we found the remarkable result that the effect of perturbations on ongoing activity predicted the way perturbations affected stimulus-evoked responses.

2.4 Changes in cluster timescale are mediated by gain modulation

What is the computational mechanism mediating the changes in cluster timescale induced by the perturbation? We argue that this phenomenon is driven by gain modulation. Using mean field theory, we can represent network attractors, defined by sets of co-activated

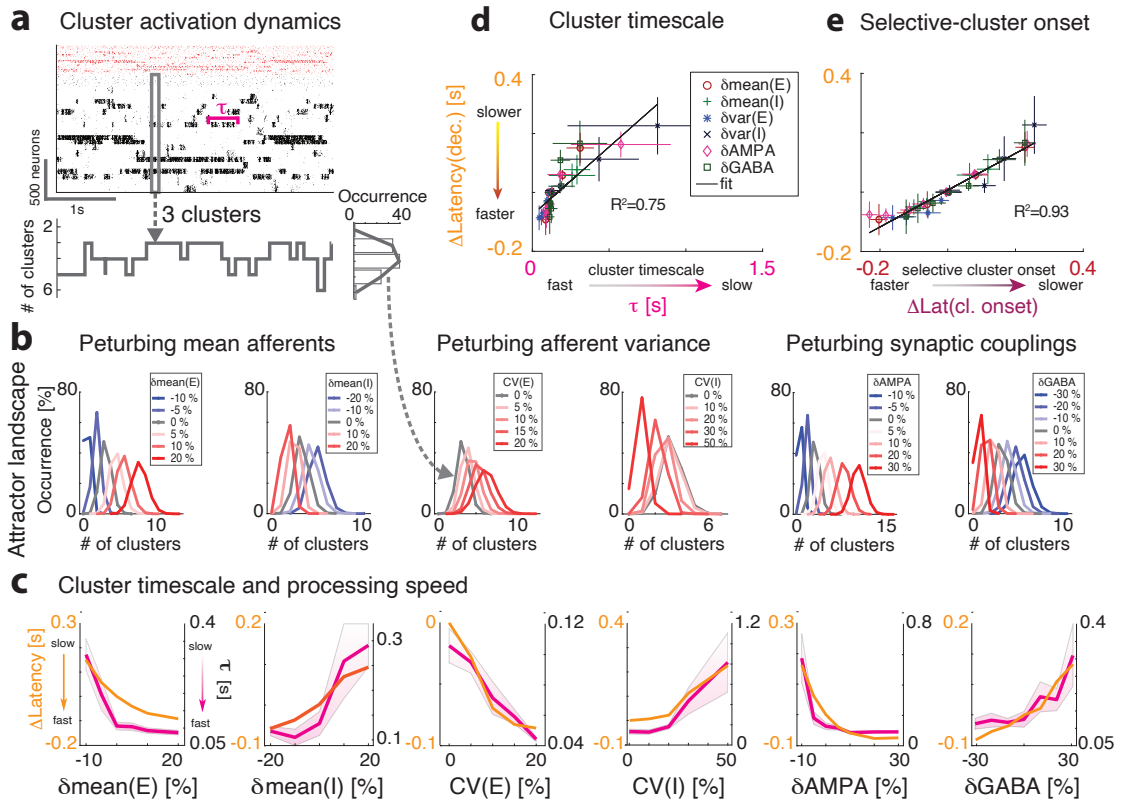

Figure 4. Perturbations modulate the cluster timescale. **a**) Top: The clustered network activity during a representative ongoing trial hops among different metastable attractors (grey box: attractor with 3 co-active clusters). Bottom: Number of co-active clusters at each time bin (right: frequency of occurrence of attractors with 2-6 co-active clusters in the representative trial). **b**) Perturbations strongly modulate the attractor landscape (color-coded curves: frequency of occurrence of network attractors for each value of the perturbation, mean occcurrence across 5 sessions). **c**): Perturbation induced changes in the cluster timescale during spontaneous periods (pink curve, mean±s.e.m. across 5 sessions) explain the changes in stimulus-decoding latency (orange curve, same as in Fig. 2d; panel **d**: linear regression, $R^2 = 0.75$). **e**) Onset latency of stimulus-selective clusters (cfr. brown vertical line in Fig. 2a-b) is modulated by the perturbation and explains the change in stimulus-decoding latency (linear regression, $R^2 = 0.93$.

clusters, as potential wells in an attractor landscape [20, 23, 25, 34, 35]. Let us illustrate this in a simplified network with two clusters (Fig. 5a and S4). Here, the attractor landscape consists of two potential wells, each well corresponding to an attractor where one cluster is active and the other is inactive. When the network activity dwells in the attractor represented by the left potential well, it may escape to the right potential well due to internally generated variability. This process will occur with a probability determined by the height Δ of the barrier separating the two wells: the higher the barrier, the less likely the transition [20, 23, 35, 36]. We found that perturbations differentially control the height of the barrier Δ separating the two attractors (Fig. S4). A mean field theory analysis showed

that the potential energy can be directly obtained from an effective transfer function for a single population (see Fig. S4, Methods and [20, 34] for details), thus establishing a direct relationship between the slope of the transfer function during ongoing periods (hereby referred to as "gain") and the barrier height Δ (Fig. 5a, S4). Because the barrier height controls the cluster activation lifetime, we thus inherited a direct relationship between gain modulation, induced by the perturbations, and changes in cluster activation timescale. In particular, perturbations inducing steeper gain will cause deeper wells and thus increase the cluster timescale, and vice versa. Using mean field theory, we demonstrated a complete classification of the differential effect of all perturbations on barrier heights and gain (Fig. S4).

We first proceed to verify these theoretical predictions, obtained in a simplified two-cluster network, in the high dimensional case of large networks with several clusters using simulations. While barrier heights and the network's attractor landscape can be exactly calculated in the simplified two-cluster network, this task is infeasible in large networks with a large number of clusters where the number of attractors is exponential in the number of clusters. On the other hand, the mean field analysis revealed that changes in barrier heights Δ are equivalent to changes in gain, and the latter can be easily estimated from spiking activity [37, 38] (Fig. 5-S4). We thus tested whether the relation between gain and timescale held in the high-dimensional case of a network with many clusters. We estimated single-cell transfer functions from their spiking activity during ongoing periods, in the absence of sensory stimuli but in the presence of different perturbations (Fig. 5b, [37, 38]). We found that perturbations strongly modulated single-cell gain in the absence of stimuli, verifying mean field theory predictions in all cases (Fig. 5c and Fig. S4). In particular, we confirmed the direct relationship between gain and cluster timescale τ, such that perturbations that decreased (increased) the gain also decreased (increased) cluster timescale (Fig. 5c, $R^2 = 0.96$, and Fig. S4). For all perturbations, gain modulations explained the observed changes in cluster timescale.

We then tested whether perturbation-induced gain modulations explained the changes in stimulus-processing speed during evoked periods, and found that the theoretical predictions (Fig. S4) were borne out in the simulations in all cases (Fig. 5d-e). Let us summarize the conclusion of our theoretical analyses. Motivated by mean field theory linking gain modulation to changes in transition rates between attractors (i.e., potential barrier heights), we found that gain modulation controls the cluster timescale during ongoing periods. We then observed that changes in cluster timescale determine changes in the onset latency of stimulus-selective clusters upon stimulus presentation. Changes in onset latency of stimulus-selective clusters explained changes in stimulus-coding latency. We thus linked gain modulation to changes in stimulus-processing speed (Fig. 5d-e).

2.5 Locomotion decreases single-cell gain and accelerates visual processing speed

Our theory predicts a link between gain modulations measured during ongoing periods and changes in stimulus-processing speed. We sought to experimentally test this prediction in freely running mice using electrophysiological recordings from primary visual cortex (V1)

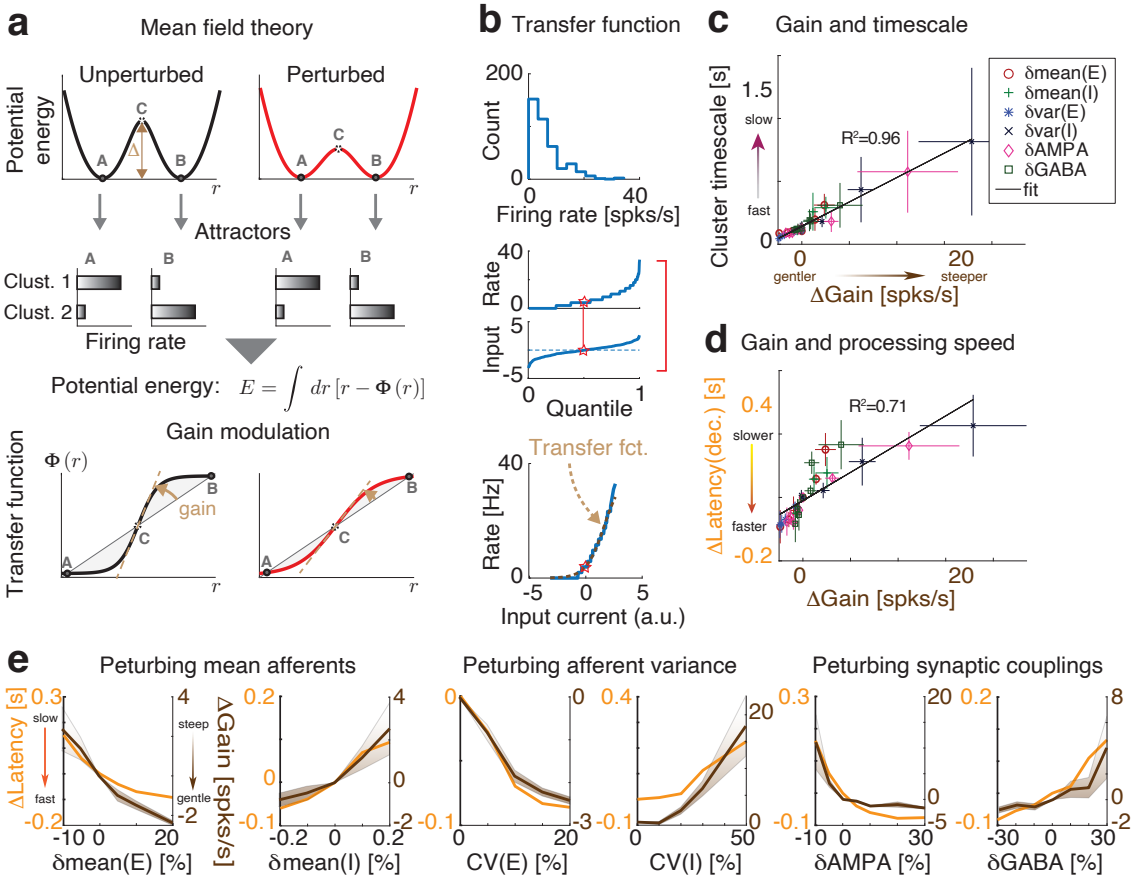

Figure 5. Linking gain modulation to cluster timescale and processing speed. **a**) Schematic of the effect of perturbations on network dynamics. Dynamics in a two-cluster network is captured by its effective potential energy (top panel). Potential wells represent two attractors where either cluster is active (A and B). A perturbations that shrinks the barrier height Δ separating the attractors induces faster transition rates between attractors and shorter cluster activation lifetime (black and red: unperturbed and perturbed conditions, respectively). Mean field theory provides a relation between potential energy and transfer function (bottom panel), thus linking cluster lifetime to neuronal gain in the absence of stimuli (dashed brown line, gain). **b**): A single-cell transfer function (bottom, empirical data in blue; sigmoidal fit in brown) can be estimated by matching a neuron's firing rate distribution during ongoing periods (top) to a gaussian distribution of input currents (center, quantile plots; red stars denotes matched median values). **c**) Perturbation-induced changes in gain (x-axis: gain change in perturbed minus unperturbed condition, mean±s.e.m. across 10 networks; color-coded markers represent different perturbations) explain changes in cluster lifetime (y-axis, linear regression, $R^2 = 0.96$) as predicted by mean field theory (**a** and Fig. S4). **d**): Perturbation-induced changes in gain during ongoing periods predict the effect of perturbations on stimulus-processing speed (linear regression, $R^2 = 0.71$). **e**) Breakdown of the relationship between Perturbation-induced gain changes and processing speed (same data as in **d**).

and 4 higher cortical visual areas (LM, AL, PM, AM; open-source neuropixels dataset available from the Allen Institute [39]). We interpreted periods where the animal was resting as akin to the "unperturbed" condition in our model, and periods where the animal

was running as the "perturbed" condition (Fig. 6a in the data). During periods of ongoing activity (in the absence of visual stimuli), we found that locomotion induced an overall increase in firing rate across visual cortical areas (Fig. 6b left), in agreement with previous studies [40–42]. More specifically, we found that locomotion induces mixed excited and inhibited responses across neurons (Fig. 6b right). Both these effects were consistent with the effect of locomotion as a perturbation inducing an increase in δvar(E) (Fig. 3b).

We then set out to test whether the theoretical link between gain modulation during ongoing periods and changes in stimulus-processing speed could explain the empirical effects of locomotion. We thus estimated the single-cell transfer functions from spiking activity during ongoing periods both when the animal was at rest and in motion. We found that locomotion strongly modulated the single-cell gain in the absence of stimuli (Fig. 6c). Specifically, we found that locomotion decreased the single-cell gain across all areas (Fig. 6d), consistent with the theoretical prediction from an increase in δvar(E) (Fig. 5e). According to our theory, this decrease in gain would predict an acceleration of stimulus-processing speed, which we then proceeded to test.

Previous studies have observed an improvement in peak decoding performance during locomotion [17], but changes in decoding latency have not been investigated. To probe the speed and accuracy of visual responses in perturbed and unperturbed conditions, we performed a cross-validated classification analysis to assess the amount of information regarding the orientation of drifting grating stimuli present in population spiking activity along the visual cortical hierarchy. Crucially, because decoding accuracy depends on sample size, we equalized number of trials between resting and running conditions. We found that trials in which the animal was running revealed both an increase in peak decoding accuracy and an anticipation of stimulus coding (shorter latency) as compared to trials where the animal was stationary (Fig. 6e-g), consistently across the whole visual hierarchy (Fig. 6g). Furthermore, the time to reach significant decoding for each area followed the anatomical hierarchy score in both unperturbed and perturbed conditions, consistent with the idea that information about the visual stimulus travels up a visual hierarchy in a feed-forward fashion (Fig. 6f, see [39]).

Given that locomotion induced an increase in firing rates in all areas (Fig. 6b), we then examined the extent to which the observed effects of locomotion (increased peak accuracy and anticipation) were merely due to the increase in firing rates. We thus matched the distribution of firing rates between running and resting (see methods and Fig. S5). We found that after rate matching the change in peak decoding accuracy decreased significantly (Fig. S6-S8). Crucially, the anticipation of stimulus processing speed induced by locomotion was still present in the rate-matched condition (Fig. S6), confirming that it was independent of changes in firing rates. The same effect was preserved in the rate-matched model simulations as well (Fig. S9). We thus concluded that the anticipation of visual processing speed induced by locomotion is consistent with a mechanism whereby locomotion decreases single-cell gain via an increase in the afferent variance δvar(E) as predicted by our theory.

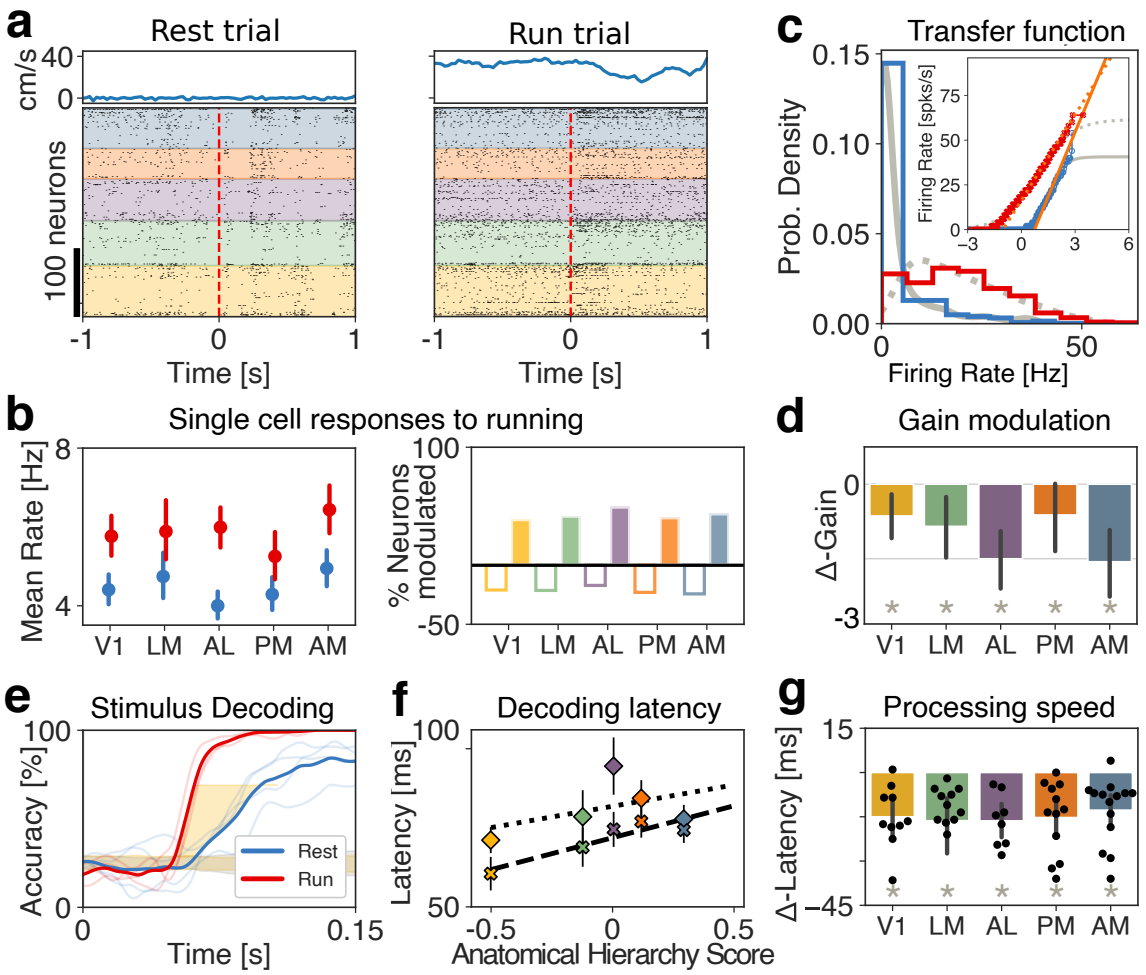

Figure 6. Effects of locomotion on visual processing. **a**) Representative raster plots from five cortical visual areas (color-coded) with population spiking activity during passive presentation of drifting gratings (dashed red line represents stimulus onset) during periods of running (right, running speed in top panels) rest (left). **b**): A representative single-cell distribution of firing rates for rest (blue) and running (red) conditions. The overlaid distributions of firing rates are obtained by passing a standard normal distribution through the sigmoidal transfer function fit shown in the inset for rest (full gray line) and running (dashed gray line). The gain for each behavioral condition (orange lines) was estimated as the slope of the sigmoidal transfer function fit at the inflection point (see Methods). **c**) Left: mean firing rate by area during rest (blue) and running (red), averaged across all periods of ongoing activity. Right: Fraction of neurons by area with significantly excited (positive bars) and inhibited (negative bars) responses to bouts of running (rank-sum test, $* = p < 0.005$). **d**): Single-cell gain modulation (Δgain=gain(running)-gain(rest)) by area across all neurons during ongoing periods (bars show 95% confidence interval; rank-sum test $* = p < 0.005$). **e**) Time course of the mean stimulus-decoding accuracy across orientations during running and rest using neurons from V1 as predictors shows the anticipation of stimulus coding in the running condition (single sessions and session average, thin and thick lines, respectively; see Methods). **f** Decoding latency (first bin above chance decoding regions in e) slows down along the anatomical hierarchy (x-axis: anatomical hierarchy score from [39]). **g**) Difference in processing speed between running and resting (average latency of decoding accuracy between 40% and 80%, yellow area in panel e) reveals running-induced coding acceleration (t-test, $* = p < 0.01$).

3 Discussion

Cortical circuits flexibly adapt their information processing capabilities to changes in environmental demands and internal state. Empirical evidence suggests that these state-dependent modulations may occur already in the sensory cortex where they may be induced by top-down pathways or neuromodulation. Here, we presented a mechanistic theory explaining how cortical stimulus-processing speed can be flexibly controlled in a state-dependent manner via gain modulation, induced by transient changes in the afferent currents or in the strength of synaptic transmission.

Our theory entails a recurrent spiking network where excitatory and inhibitory neurons are arranged in clusters, generating metastable activity in the form of transient activation of subsets of clusters. We showed that gain modulation controls the timescale of metastable activity and thus the network's information-processing speed and reaction times. In particular, our theory predicted that perturbations that decrease (increase) the intrinsic single-cell gain during ongoing periods accelerate (slow down) the latency of stimulus responses.

We tested this prediction by examining the effect of locomotion on visual processing in freely running mice. We found that locomotion reduced the intrinsic single-cell gain during ongoing periods, thus accelerating stimulus-coding speed across the visual cortical hierarchy. Our theory suggests that the observed effects of locomotion are consistent with a perturbation that increases the spatial variance of the afferent currents to the local excitatory population. These results establish a new theory of state-dependent adaptation of cortical responses via gain modulation, unifying the effect of different pathways under a shared computational mechanism.

3.1 Metastable activity in cortical circuits

The crucial dynamical feature of our model is its metastable activity, whereby single-trial ensemble spike trains unfold through sequences of metastable states. State are long-lasting, with abrupt transitions between consecutive states. Metastable activity has been ubiquitously observed in a variety of cortical and subcortical areas, across species and tasks [43–51]. Metastable activity can be used to predict behavior and was implicated as a neural substrate of cognitive function, such as attention [45], expectation [20], and decision making [38, 48, 50]. Metastable activity was observed also during ongoing periods, in the absence of sensory stimulation, suggesting that it may be an intrinsic dynamical regime of cortical circuits [25, 45]. Here, we showed how cortical circuits can flexibly adjust their performance and information-processing speed via modulations of their metastable dynamics.

Metastable activity may naturally arise in circuits where multiple stable states, or attractors, are destabilized by external perturbations [52] or intrinsically generated variability [20, 23–25, 27, 28, 38]. Biologically plausible models of metastable dynamics have been proposed in terms of recurrent spiking networks where neurons are arranged in clusters, reflecting the empirically observed assemblies of functionally correlated neurons [29–32]. Clustered network models of metastable dynamics provide a parsimonious explanation of several physiological observations such as stimulus-induced reductions of trial-to-trial variability [24, 25, 27, 53, 54], of firing rate multistability [25], and of neural dimensionality

[26]. Our results extend the biological plausibility of clustered networks by showing that they capture other ubiquitous features of cortical dynamics: they operate in the inhibition stabilized regime [55–57]; they naturally give rise to lognormal distribution of firing rates [58–61]. This class of models thus provide a biologically plausible, mechanistic link between connectivity, dynamics, and information-processing.

3.2 Linking metastable activity to flexible cognitive function via gain modulation

Recent studies have shown that cortical circuits may implement a variety of flexible cognitive computations by modulating the timescale of their intrinsic metastable dynamics [20, 45, 48, 51]. Our results establish a comprehensive framework to investigate the extent of this hypothesis. We propose that gain modulation is the neural mechanism underlying flexible state-dependent cortical computation. Specifically, we showed that gain modulation controls the timescale of metastable dynamics, which, in turn, determines the network's information-processing speed.

3.3 Alternative models of gain modulation

Previous studies have suggested gain modulation as a mechanism to sharpen single-cell tuning curves without affecting selectivity [62, 63], potentially mediating attention [64–66]. In those studies, gain modulation was defined as change in the single-neuron response function to stimuli of increasing contrast. Here, we have taken a different approach and defined gain as the slope of the intrinsic neuronal current-to-rate function during ongoing periods (i.e., in the absence of stimuli, see also [20, 63, 67]), as opposed to the contrast response function. We have classified mechanisms of gain modulation which act by changing the mean or spatial variance *across neurons* of the cell-type specific afferent currents to the local cortical circuit, where we modeled afferent currents as constant biases; or by changing the recurrent couplings. The rationale for our choice was to investigate the effects on internally generated variability in a network whose dynamics were entirely deterministic. Alternatively, one could model external currents as time-dependent inputs with fast noise, such as Poisson processes or colored noise. In that case, changes in background noise due to barrages of synaptic inputs are capable of inducing gain modulation as well [63, 67]. Previous work compared these different kinds of perturbations (Poisson noise or afferent spatial variance) in the case of the perturbation $\delta\mathrm{var(E)}$ [20], showing they may lead to similar outcomes.

3.4 Physiological mechanisms of gain modulation

Several different physiological pathways can modulate the gain of the intrinsic neuronal transfer function, including neuromodulation, top-down and cortico-cortical interactions. Gain modulation can also be induced artificially by means of optogenetic or pharmacological manipulations. The perturbations investigated in our model may be related to different pathways and implicated in various types of cognitive function.

3.4.1 Neuromodulation

Neuromodulatory pathways strongly affect sensory processing in cortical circuits by changing cell-type specific afferent currents to the circuit, in some cases controlling their dynamical regime [15]. Our theory may be applicable to explain the effects of cholinergic and serotonergic activation on sensory cortex.

Cholinergic pathways, modulating ionic currents in pyramidal cells [68], can control cortical states and mediate the effects of arousal and locomotion. Artificial stimulation of cholinergic pathways was shown to improve sensory coding in visual [8, 69] and barrel cortex [70]. Cholinergic stimulation alone in the absence of sensory stimuli was shown to induce mixed responses with different neural populations increasing or decreasing their spiking activity [69]. Our theory shows that these combined experimental observations (coding improvement and mixed firing rate changes) are consistent with a mechanism whereby cholinergic activation induces an increase in δvar(E) afferents to sensory cortex, inducing an acceleration of sensory processing (Fig. 2e and 3b).

Activation of serotonergic pathways by stimulation of dorsal raphe serotonergic neurons or local iontophoresis was shown to transiently degrade stimulus coding in sensory cortex, decreasing responses to mechanosensory stimuli [71] and increasing the latency of the first spike evoked by auditory stimuli [72]. Serotonergic stimulation was shown to decrease firing rates in the olfactory cortex [73], inferior colliculus [72], and primary visual cortex [74]. Our theory shows that these experimental observations (coding degradation and decreased firing rates) are consistent with two alternative mechanisms (Fig. 2d and 3b): either an increase in the afferent currents to I populations (i.e., δmean(I)> 0) implementing the paradoxical inhibition effect [33]; or a decrease in the afferents to E populations (i.e., δmean(E)< 0). Future experiments could test between these two alternatives.

3.4.2 Top-down projections

A prominent feature of sensory cortex is the integration of feedforward and cortico-cortical feedback pathways at each stage of sensory processing [75]. In particular, top-down projections from higher cortical areas to sensory cortex are known to modulate the speed and accuracy of sensory processing [20]. Our theory may explain the effects of activating several cortico-cortical pathways.

Activation of feedback axons from motor cortex (M1) to somatosensory cortex (S1) was shown to increase activity in S1 during whisking [76] and led to faster and more accurate responses to whisker stimulation [77]. Suppression of the same pathway induced slower S1 responses to whisking in awake mice. Our theory shows that the effect of these cortico-cortical perturbations is consistent with an increase in the mean afferent currents to E populations in S1 (i.e., the δmean(E) perturbation in Fig. 2d), leading to higher firing rates and faster processing speed.

Expectation and arousal are known to strongly modulate neural activity in sensory cortices [78]. Expected stimuli are processed faster and more accurately than unexpected stimuli both in auditory [5] and gustatory cortex [6]. Experimental evidence shows that the anticipation of sensory processing induced by expectation is mediated by top-down

projections from the amygdala to the gustatory cortex [6], whose activation elicits complex excited and inhibited responses in both pyramidal and inhibitory cells in the gustatory cortex [6, 79]. Our theory suggests that these top-down projections may operate by inducing an increase in the spatial variance of the afferent currents to the E population (δvar(E) in Fig. 2d), extending previous results [20] to networks including inhibitory clusters.

In attentional tasks, distractors slow down reaction times [2, 3], a behavioral effect that may be mediated by changes in the speed and accuracy of sensory processing in cortical circuits [4]. The presence of distracting stimuli within a neurons receptive field suppresses its responses to the preferred stimulus [80]. The underlying mechanism may recruit lateral inhibition onto the local cortical circuit [81, 82]. Our theory shows that this mechanism is consistent with a modulation of the afferents to local I populations, mediated by either an increase in δmean(I) or δvar(I) in Fig. 2d. It would be interesting to discriminate between these two perturbations with future experiments.

3.4.3 Optogenetic and pharmacological manipulations

Our theory may shed light on the effects of manipulation experiments. Optogenetic activation (inactivation) of specific E or I cells [83, 84] has been modeled as an increase (decrease) of the afferent currents to those cells [56, 85, 86]. However, protein expression may not be complete across all cells of the targeted population, and even in the case of complete expression across the targeted population, different cells may be more or less sensitive to laser stimulation. Thus the effect of optogenetic stimulation on the targeted population may then be more accurately modeled by a concurrent change in both mean and variance of the targeted cell-type specific afferents (e.g., δmean(E) and δvar(E) for E populations; δmean(I) and δvar(I) for I populations). Recent studies showed that, while a homogeneous stimulation of all I cell types simultaneously can be captured by a model of E-I recurrently coupled neurons (as in our model), partial activation of specific inhibitory cell-types may induced more complex responses [56, 84, 86–88]. We plan to revisit this issue in the future.

Our theory may also be applicable to the effects of pharmacological manipulations of different synaptic receptors. In particular, the effects of combined local injection of AMPA/kainate and NMDA receptor antagonists (agonists) may be recapitulated by a decrease (increase) in δAMPA, which correspondingly perturb the value of J_{IE}, J_{EE} couplings (Fig. 2d). Similarly, the effects of local injection of GABA receptor antagonists (agonists) may be recapitulated by a decrease (increase) in δGABA, which correspondingly perturb the value of J_{EI}, J_{II} couplings.

3.5 Locomotion and gain modulation

Locomotion has been shown to modulate visually evoked activity [40] and is sufficient in driving activity in mouse V1 [13, 89]. Our results were consistent with previous studies in showing that locomotion affects the activity of neurons in the visual cortical hierarchy during both ongoing and stimulus-evoked activity. We found that locomotion in the absence of sensory stimuli induces an average increase in firing rates. At the single-cell level we reported a complex mix of excited and inhibited responses in both E and I cells, also consistent with previous results [9, 42]. Crucially, we uncovered that locomotion decreased the single-cell

Model parameters for clustered network simulations		
Parameter	Description	Value
j_{EE}	mean E-to-E synaptic weights $\times\sqrt{N}$	0.6 mV
j_{IE}	mean E-to-I synaptic weights $\times\sqrt{N}$	0.6 mV
j_{EI}	mean I-to-E synaptic weights $\times\sqrt{N}$	1.9 mV
j_{II}	mean I-to-I synaptic weights $\times\sqrt{N}$	3.8 mV
j_{E0}	mean I-to-I synaptic weights $\times\sqrt{N}$	2.6 mV
j_{I0}	mean I-to-I synaptic weights $\times\sqrt{N}$	2.3 mV
δ	standard deviation of the synaptic weight distribution	20%
J_{EE}^{+}	Potentiated intra-cluster E-to-E weight factor	14
J_{II}^{+}	Potentiated intra-cluster I-to-I weight factor	5
g_{EI}	Potentiation parameter for intra-cluster I-to-E weights	10
g_{IE}	Potentiation parameter for intra-cluster E-to-I weights	8
r_{ext}	Average baseline afferent rate to E and I neurons	5 spks/s
V_E^{thr}	E-neuron threshold potential	1.43 mV
V_I^{thr}	I-neuron threshold potential	0.74 mV
V^{reset}	E- and I-neuron reset potential	0 mV
τ_m	E- and I-neuron membrane time constant	20 ms
τ_{refr}	E- and I-neuron absolute refractory period	5 ms
τ_s	E- and I-neuron synaptic time constant	5 ms

Table 1. Parameters for the clustered network used in the simulations.

gain in the absence of visual stimuli across the board in the visual hierarchy (Fig. 6). Our theory predicted that the observed decrease in gain would lead to an acceleration of visual processing during locomotion (Fig. 2c), which we confirmed in the data (Fig. 6). Such acceleration of processing speed did not depend on the locomotion-induced changes in firing rates and was still present even after matching the firing rate distributions between running and rest conditions (Fig. S5). Our results (increased firing rates with mixed excited and inhibited responses, and faster visual processing) suggest that the effect of locomotion may be mediated by a increase in the spatial variance of the afferent current to the E populations (δvar(E) perturbation) [9, 40, 90]. Concretely, gain modulation may the combined effect of activating neuromodulatory pathways such as cholinergic [9] and noradrenergic [91] inputs.

4 Methods

4.1 Spiking network model

Architecture. We modeled the local cortical circuit as a network of $N = 2000$ excitatory (E) and inhibitory (I) neurons (with relative fraction $n_E = 80\%$ and $n_I = 20\%$) with random recurrent connectivity (Fig. 1). Connection probabilities were $p_{EE} = 0.2$ and $p_{EI} = p_{IE} = p_{II} = 0.5$. Nonzero synaptic weights from pre-synaptic neuron j to post-synaptic neuron i were $J_{ij} = j_{ij}/\sqrt{N}$, with j_{ij} sampled from a gaussian distribution with mean $j_{\alpha\beta}$, for $\alpha, \beta = E, I$, and standard deviation δ^2. E and I neurons were arranged in

Model parameters for the reduced two-cluster network		
Parameter	Description	Value
j_{EE}	mean E-to-E synaptic weights $\times\sqrt{N}$	0.8 mV
j_{EI}	mean I-to-E synaptic weights $\times\sqrt{N}$	10.6 mV
j_{IE}	mean E-to-I synaptic weights $\times\sqrt{N}$	2.5 mV
j_{II}	mean I-to-I synaptic weights $\times\sqrt{N}$	9.7 mV
j_{E0}	mean I-to-I synaptic weights $\times\sqrt{N}$	14.5 mV
j_{I0}	mean I-to-I synaptic weights $\times\sqrt{N}$	12.9 mV
J^+_{EE}	Potentiated intra-cluster E-to-E weight factor	9
r^{ext}	Average baseline afferent rate to E and I neurons	7 spk/s
V^{thr}_E	E-neuron threshold potential	4.6 mV
V^{thr}_I	I-neuron threshold potential	8.7 mV
τ_s	E- and I-neuron synaptic time constant	4 ms
n_{bgr}	Fraction of background E neurons	65%

Table 2. Parameters for the simplified two-cluster network used for the mean-field theory analysis (the remaining parameters are in Table 1.

p clusters. E clusters had heterogeneous sizes drawn from a gaussian distribution with a mean of $N_E^{clust} = 80$ E-neurons and 20% standard deviation. The number of clusters was then determined as $p = \text{round}(n_E N(1 - n_{bgr})/N_E^{clust})$, where $n_{bgr} = 0.1$ is the fraction of background neurons in each population, i.e., not belonging to any cluster. I clusters had equal size $N_I^{clust} = \text{round}(n_I N(1 - n_{bgr}/p)$. Synaptic weights for within-cluster neurons where potentiated by a ratio factor $J^+_{\alpha\beta}$. Synaptic weights between neurons belonging to different clusters were depressed by a factor $J^-_{\alpha\beta}$. Specifically, we chose the following scaling: $J^+_{EI} = p/(1 + (p-1)/g_{EI})$, $J^+_{IE} = p/(1 + (p-1)/g_{IE})$, $J^-_{EI} = J^+_{EI}/g_{EI}$, $J^-_{IE} = J^+_{IE}/g_{IE}$ and $J^-_{\alpha\alpha} = 1 - \gamma(J^+_{\alpha\alpha} - 1)$ for $\alpha = E, I$, with $\gamma = f(2 - f(p+1))^{-1}$, where $f = (1 - n_{bgr})/p$ is the fraction of E neurons in each cluster. Within-cluster E-to-E synaptic weights were further multiplied by cluster-specific factor equal to the ratio between the average cluster size N_E^{clust} and the size of each cluster, so that larger clusters had smaller within-cluster couplings. We chose network parameters so that the cluster timescale was 100 ms, as observed in cortical circuits [20, 25, 44]. Parameter values are in Table 1.

Neuronal dynamics. We modeled spiking neurons as current-based leaky-integrate-and-fire (LIF) neurons whose membrane potential V evolved according to the dynamical equation

$$\frac{dV}{dt} = \frac{V}{\tau_m} + I_{rec} + I_{ext} \ ,$$

where τ_m is the membrane time constant. Input currents included a contribution I_{rec} coming from the other recurrently connected neurons in the local circuit and an external current $I_{ext} = I_0 + I_{stim} + I_{pert}$ (units of mV s^{-1}). The first term $I_0 = N_{ext} J_{\alpha 0} r_{ext}$ (for $\alpha = E, I$) is a constant term representing input to the E or I neuron from other brain areas and $N_{ext} = n_E N p_{EE}$; while I_{stim} and I_{pert} represent the incoming sensory stimulus or the various types of perturbation (see Stimuli and perturbations below). When V hits

threshold V_α^{thr} (for $\alpha = E, I$), a spike is emitted and V is then held at the reset value V^{reset} for a refractory period τ_{refr}. We chose the thresholds so that the homogeneous network (i.e.,where all $J_{\alpha\beta}^{\pm} = 1$) was in a balanced state with average spiking activity at rates $(r_E, r_I) = (2, 5)$ spks/s [20, 22]. Recurrent synapses evolved according to the following equation

$$\tau_{syn} \frac{dI_{rec}}{dt} = -I_{rec} + \sum_{j=1}^{N} J_{ij} \sum_k \delta(t - t_k) \ ,$$

where τ_s is the synaptic time constant, J_{ij} are the recurrent couplings and t_k is the time of the k-th spike from the j-th presynaptic neuron. Parameter values are in Table 1.

Stimuli and perturbations. We considered two classes of inputs: sensory stimuli and perturbations. In the "evoked" condition (Fig. 2a), We presented the network 4 sensory stimuli, modeled as changes in the afferent currents targeting 50% of E-neurons in stimulus-selective clusters; each E-cluster had a 50% probability of being selective to a sensory stimulus (mixed selectivity). I-clusters were not stimulus-selective. In both the unperturbed and the perturbed stimulus-evoked conditions, stimulus onset occurred at time $t = 0$ and each stimulus was represented by an afferent current $I_{stim}(t) = I_{ext} r_{stim}(t)$, where $r_{stim}(t)$ is a linearly ramping increase reaching a value $r_{max} = 20\%$ above baseline at $t = 1$. We considered several kinds of perturbations. In the perturbed stimulus-evoked condition (Fig. 2a, right panel), perturbation onset occurred at time $t = -0.5$ and lasted until the end of the stimulus presentation at $t = 1$ with a constant time course. We also presented perturbations in the absence of sensory stimuli ("ongoing" condition, Fig. 4); in that condition, the perturbation was constant and lasted for the whole duration of the trial (5s). Finally, when assessing single-cell responses to perturbations, we modeled the perturbation time course as a double exponential with rise and decay times $[0.1, 1]$s (Fig. 3). In all conditions, perturbations were defined as follows:

- δmean(E), δmean(I): A constant offset $I_{pert} = zI_0$ in the mean afferent currents was added to all neurons in either E or I populations, respectively, expressed as a fraction of the baseline value I_0 (see Neuronal dynamics above), where $z \in [-0.1, 0.2]$ for E neurons and $z \in [-0.2, 0.2]$ for I neurons.

- δvar(E), δvar(I): For each E or I neuron, respectively, the perturbation was a constant offset $I_{pert} = zI_0$, where z is a gaussian random variable with zero mean and standard deviation σ_z. We chose $\sigma_z \in [0, 0.2]$ for E neurons and $\sigma_z \in [0, 0.5]$ for I neurons. This perturbation did not change the mean afferent current but only its spatial variance across the E or I population, respectively.

- δAMPA: A constant change in the mean $j_{\alpha E} \to (1+z) j_{\alpha E}$ synaptic couplings (for $\alpha = E, I$), representing a modulation of glutamatergic synapses. We chose $z \in [-0.1, 0.2]$.

- δGABA: A constant change in the mean $j_{\alpha I} \to (1+z) j_{\alpha I}$ synaptic couplings (for $\alpha = E, I$), representing a modulation of GABAergic synapses. We chose $z \in [-0.2, 0.2]$.

The range of the perturbations were chosen so that the network still produced metastable dynamics for all values.

Inhibition stabilization. We simulated a stimulation protocol used in experiments to test inhibition stabilization (Fig. 2e). This protocol is identical to the δmean(I) perturbation during ongoing periods, where the perturbation targeted all I neurons with an external current $I_{pert} = zI_0$ applied for the whole length of 5s intervals, with $z \in [0, 1.2]$ and 40 trials per network and 10 networks for each value of the perturbation.

Simulations. All data analyses, model simulations, and mean field theory calculations were performed using custom software written in MATLAB, C and Python. Simulations in the stimulus-evoked conditions (both perturbed and unperturbed) comprised 10 realizations of each network (each network with different realization of synaptic weights), with 20 trials for each of the 4 stimuli. Simulations in the ongoing condition comprised 10 different realization of each network, with 40 trials per perturbation. Each network was initialized with random synaptic weights and simulated with random initial conditions in each trial. Sample sizes were similar to those reported in previous publications [20, 25, 26]. Dynamical equations for the leaky-integrate-and-fire neurons were integrated with the Euler method with a 0.1ms step.

4.2 Mean field theory

We performed a mean field analysis of a simplified two-cluster network for leaky-integrate-and-fire neurons with exponential synapses, comprising $p+2$ populations for $p = 2$ [20, 22]: the first p representing the two E clusters, the last two representing the background E and the I population. The infinitesimal mean μ_n and variance σ_n^2 of the postsynaptic currents are:

$$\begin{aligned}
\mu_n &= \tau_m\sqrt{N}\left[n_E p_{EE} j_{EE}\left(fJ^+_{EE} r_n + J^-_{EE}(\sum_{l=1}^{p-1} r_l + (1-pf)r_E^{bgr}) + \frac{j_{E0}}{j_{EE}} r_{ext}\right) - n_I p_{EI} j_{EI} r_I\right] , \\
\mu_{bgr} &= \tau_m\sqrt{N}\left[n_E p_{EE} j_{EE}\left(J^-_{EE}\sum_{l=1}^{p} r_l + (1-pf)r_E^{bgr} + \frac{j_{E0}}{j_{EE}} r_{ext}\right) - n_I p_{EI} j_{EI} r_I\right] , \\
\mu_I &= \tau_m\sqrt{N}\left[n_E p_{IE} j_{IE}\left(f\sum_{l=1}^{p} r_l + (1-pf)r_E^{bgr}\right) - n_I p_{II}(j_{II} r_I + j_{I0} r_{ext})\right] ,
\end{aligned} \tag{4.1}$$

$$\begin{aligned}
\sigma_n^2 &= \tau_m\sqrt{N}[n_E p_{EE} j_{EE}^2\left(f(J^+_{EE})^2 r_n + (J^-_{EE})^2(\sum_{l=1}^{p-1} r_l + (1-pf)r_E^{bgr})\right) - n_I p_{EI} j_{EI}^2 r_I] , \\
\sigma_{bgr}^2 &= \tau_m\sqrt{N}\left[n_E p_{EE} j_{EE}^2\left((J^-_{EE})^2\sum_{l=1}^{p} r_l + (1-pf)r_E^{bgr}\right) - n_I p_{EI} j_{EI}^2 r_I\right] , \\
\sigma_I &= \tau_m\sqrt{N}\left[n_E p_{IE} j_{IE}^2\left(f\sum_{l=1}^{p} r_l + (1-pf)r_E^{bgr}\right) - n_I p_{II} j_{II}^2 r_I\right] ,
\end{aligned} \tag{4.2}$$

where $r_n, r_l = 1, \ldots, p$ are the firing rates in the p E-clusters; r_E^{bgr}, r_I, r_{ext} are the firing rates in the background E population, in the I population, and in the external current. Other parameters are described in Architecture and in Table 2. The network attractors satisfy

the self-consistent fixed point equations:

$$r_l = F_l[\mu_l(\mathbf{r}), \sigma_l^2(\mathbf{r})] \ , \tag{4.3}$$

where $\mathbf{r} = (r_1, \ldots, r_p, r_{bgr}, r_I)$ and $l = 1, \ldots, p, bgr, I$, and F_l is the current-to-rate transfer function for each population, which depend on the condition. In the absence of perturbations, all populations have the LIF transfer function

$$F_l(\mu_l, \sigma_l) = \left(\tau_{refr} + \tau_m \sqrt{\pi} \int_{H_l}^{\Theta_l} e^{u^2} [1 + (u)] \right)^{-1} \ , \tag{4.4}$$

where $H_l = (V^{reset} - \mu_l)/\sigma_l + ak$ and $\Theta_l = (V_l^{thr} - \mu_l)/\sigma_l + ak$. $k = \sqrt{\tau_s/\tau_m}$ and $a = |\zeta(1/2)|/\sqrt{2}$ are terms accounting for the synaptic dynamics [92]. The perturbations δvar(E) and δvar(I) induced an effective population transfer function F^{eff} on the E and I populations, respectively, given by [20]:

$$F_\alpha^{pert}(\mu_\alpha, \sigma_\alpha) = \int Dz F_\alpha(\mu_\alpha + z\sigma_z \mu_\alpha^{ext}, \sigma_\alpha^2) \ , \tag{4.5}$$

where $\alpha = E, I$ and $Dz = \exp(-z^2/2/\sqrt{2\pi})$ is a gaussian measure of zero mean and unit variance, $\mu_\alpha^{ext} = \tau_m \sqrt{N} n_\alpha p_{\alpha 0} j_{\alpha 0} r_{ext}$ is the external current and σ_z is the standard deviation of the perturbation with respect to baseline, denoted CV(E) and CV(I) in Fig. S4 and in the Results. Stability of the fixed point equation 4.3 was defined with respect to the approximate linearized dynamics of the instantaneous mean m_l and variance s_l^2 of the input currents:[20, 25]

$$\tau_s \frac{dm_l}{dt} = -m_l + \mu_l(r_l) \ ; \quad \tau_s \frac{ds_l^2}{2dt} = -s_l^2 + \sigma_l^2(r_l) \ ; \quad r_l = F_l(m_l(\mathbf{r}), s_l^2(\mathbf{r})) \ , \tag{4.6}$$

where μ_l, σ_l^2 are defined in 4.1-4.2 and F_l represents the appropriate transfer function 4.4 or 4.5. Fixed point stability required that the stability matrix

$$S_{lm} = \frac{1}{\tau_s} \left(\frac{\partial F_l(\mu_l, \sigma_l^2)}{\partial r_m} - \frac{\partial F_l(\mu_l, \sigma_l^2)}{\partial \sigma_l^2} \frac{\partial \sigma_l^2(\mathbf{r})}{\partial r_m} - \delta_{lm} \right) \ , \tag{4.7}$$

was negative definite. The full mean field theory described above was used for the comprehensive analysis of Fig. S4. For the schematic of Fig. 5a, we replaced the LIF transfer function 4.4 with the simpler function $\tilde{F}(\mu_E) = 0.5(1 + \tanh(\mu_E))$ and the δvar(E) perturbation effect was then modeled as $\tilde{F}^{eff}(\mu) = \int Dz \tilde{F}(\mu_E + z\sigma_z \mu_{ext})$.

Effective mean field theory for a reduced network. To calculate the potential energy barrier separating the two network attractors in the reduced two-cluster network, we used the effective mean field theory developed in [20, 34]. The idea is to first estimate the force acting on neural configurations with cluster firing rates $\mathbf{r} = [\tilde{r}_1, \tilde{r}_2]$ outside the fixed points (4.3), then project the two-dimensional system onto a one-dimensional trajectory along which the force can be integrated to give an effective potential E (Fig. S4). In the first step, we start from the full mean field equations for the $P = p + 2$ populations in 4.3, and obtain an effective description of the dynamics for q populations "in focus" describing E

clusters ($q = 2$ in our case) by integrating out the remaining $P - q$ out-of-focus populations describing the background E neurons and the I neurons ($P - q = 2$ in our case). Given a fixed value $\tilde{\mathbf{r}} = [\tilde{r}_1, \ldots, \tilde{r}_q]$ for the q in-focus populations, one obtains the stable fixed point firing rates $\mathbf{r}' = [r'_{q+1}, \ldots, r'_P]$ of the out-of-focus populations by solving their mean field equations

$$r'_\beta(\tilde{\mathbf{r}}) = F_\beta[\mu_\beta(\tilde{\mathbf{r}}, \mathbf{r}'), \sigma^2_\beta(\tilde{\mathbf{r}}, \mathbf{r}')] \ , \tag{4.8}$$

for $\beta = q+1, \ldots, P$, as function of the in-focus populations $\tilde{\mathbf{r}}$, where stability is calculated with respect to the condition (4.7) for the reduced $(q+1, \ldots, P)$ out-of-focus populations at fixed values of the in-focus rates $\tilde{\mathbf{r}}$. One then obtains a relation between the input $\tilde{\mathbf{r}}$ and output values $\tilde{\mathbf{r}}^{out}$ of the in-focus populations by inserting the fixed point rates of the out-of-focus populations calculated in (4.8):

$$r^{out}_\alpha(\tilde{\mathbf{r}}) = F_\alpha[\mu_\alpha(\tilde{\mathbf{r}}, \mathbf{r}'(\tilde{\mathbf{r}})), \sigma^2_\alpha(\tilde{\mathbf{r}}, \mathbf{r}'(\tilde{\mathbf{r}}))] \ , \tag{4.9}$$

for $\alpha = 1, \ldots, q$. The original fixed points are $\tilde{\mathbf{r}}^*$ such that $\tilde{r}^*_\alpha = r^{out}_\alpha(\tilde{\mathbf{r}^*})$.

Potential energy barriers and transfer function gain. In a reduced network with two in-focus populations $[\tilde{r}_1, \tilde{r}_2]$ corresponding to the two E clusters, one can visualize Eq. (4.9) as a two-dimensional force vector $\tilde{\mathbf{r}} - \mathbf{r}^{out}(\tilde{\mathbf{r}})$ at each point in the two-dimensional firing rate space $\tilde{\mathbf{r}}$. The force vanishes at the stable fixed points A and B and at the unstable fixed point C between them (Fig. S4). One can further reduce the system to one dimension by approximating its dynamics along the trajectory between A and B as [34]:

$$\tau_s \frac{d\tilde{r}}{dt} = -\tilde{r} + r^{out}(\tilde{r}) \ , \tag{4.10}$$

where $y = r^{out}(\tilde{r})$ represents an effective transfer function and $\tilde{r} - r^{out}(\tilde{r})$ an effective force. We estimated the gain g of the effective transfer function as $g = 1 - \frac{r^{out}(\tilde{r}_{min}) - r^{out}(\tilde{r}_{min})}{\tilde{r}_{min} - \tilde{r}_{max}}$, where $\tilde{r}_{min}$ and $\tilde{r}_{max}$ represent, respectively, the minimum and maximum of the force (see Fig. S4). From the one-dimensional dynamics (4.10) one can define a potential energy via $\frac{\partial E(\tilde{r})}{\partial r} = \tilde{r} - r^{out}(\tilde{r})$. The energy minima represent the stable fixed points A and B and the saddle point C between them represents the potential energy barrier separating the two attractors. The height Δ of the potential energy barrier is then given by

$$\Delta = \int_A^C d\tilde{r}[\tilde{r} - r^{out}(\tilde{r})] \ , \tag{4.11}$$

which can be visualized as the area of the curve between the effective transfer function and the diagonal line (see Fig. 5).

4.3 Experimental data

We tested our model predictions using the open-source dataset of neuropixel recordings from the Allen Institute for Brain Science [39]. We focused our analysis on experiments where drifting gratings were presented at four directions (0°, 45°, 90°, 135°) and one temporal frequency (2 Hz). Out of the 54 sessions provided, only 7 sessions had enough trials per behavioral condition to perform our decoding analysis. V1 was collected in 5 of these 7

sessions, with a median value of 75 neurons per session (LM: 6 sessions, 47 neurons; AL: 5 sessions, 61 neurons; PM: 6 sessions, 55; AM: 7 sessions, 48 neurons). We matched the number and duration of trials across condition and orientation and combined trials from the drifting gratings repeat stimulus set, and drifting grating contrast stimulus set. To do this, we combined trials with low-contrast gratings (0.08, 0.1, 0.13, 0.2; see Fig. S8) and trials with high-contrast gratings (0.6, 0.8, 1; see Fig. S7) into separate trial types to perform the decoding analysis, and analyzed the interval $[0, 0.5]$ seconds aligned to stimulus onset.

For evoked activity, running trials were classified as those where the animal was running faster than 3 cm/s for the first 0.5 seconds of stimulus presentation. During ongoing activity, behavioral periods were broken up into windows of 1 second. Periods of running or rest were classified as such if 10 seconds had elapsed without a behavioral change. Blocks of ongoing activity were sorted and used based on the length of the behavior. Out of the 54 sessions provided, 14 sessions had enough time per behavioral condition (minimum of 2 minutes) to estimate single-cell transfer functions. Only neurons with a mean firing rate during ongoing activity greater than 5Hz were included in the gain analysis (2119 out of 4365 total neurons).

4.4 Stimulus decoding

For both the simulations and data, a multi-class decoder was trained to discriminate between four stimuli from single-trial population activity vectors in a given time bin [93]. To create a timecourse of decoding accuracy, we used a sliding window of 100ms (200ms) in the data (model), which was moved forward in 2ms (20ms) intervals in the data (model). Trials were split into training and test data-sets in a stratified 5-fold cross-validated manner, ensuring equal proportions of trials per orientation in both data-sets. In the model, a leave-2-out cross-validation was performed. To calculate the significance of the decoding accuracy, an iterative shuffle procedure was performed on each fold of the cross-validation. On each shuffle, the training labels were shuffled and the classifer accuracy was predicted on the unshuffled test data-set. This shuffle was performed 100 times to create a shuffle distribution to rank the actual decoding accuracy from the unshuffled decoder against and to determine when the mean decoding accuracy had increased above chance. This time point is what we referred to as the latency of stimulus decoding. To account for the speed of stimulus decoding (the slope of the decoding curve), we defined the Δ-Latency between running and rest as the average time between the two averaged decoding curves from 40% up to 80% of the max decoding value at rest.

4.5 Firing rate distribution match

To control for increases of firing rate due to locomotion (see Fig. 6c), we matched the distributions of population counts across the trials used for decoding in both behavioral conditions. This procedure was done independently for each sliding window of time along the decoding time course. Within each window, the spikes from all neurons were summed to get a population spike count per trial. A log-normal distribution was fit to the population counts across trials for rest and running before the distribution match (Fig S5a left). We sorted the distributions for rest and running in descending order, randomly removing

spikes from trials in the running distribution to match the corresponding trials in the rest distribution (Fig S5a right). By doing this, we only removed the number of spikes necessary to match the running distribution to rest distribution. For example, trials where the rest distribution had a larger population count, no spikes were removed from either distribution. Given we performed this procedure at the population level rather than per neuron, we checked the change in PSTH between running and rest conditions before and after distribution matching (Fig S5b). This procedure was also performed on the simulated data (Fig. S9).

4.6 Single-cell gain

To infer the single-cell transfer function in simulations and data, we followed the method originally described in [38] (see also [37, 94] for a trial-averaged version). We estimated the transfer function on ongoing periods when no sensory stimulus was present. Briefly, the transfer function of a neuron was calculated by mapping the quantiles of a standard gaussian distribution of input currents to the quantiles of the empirical firing rate distribution during ongoing periods (see 5b). We then fit this transfer function with a sigmoidal function. The max firing rate of the neuron in the sigmoidal fit was bounded to be no larger than 1.5 times that of the empirical max firing rate, to ensure realistic fits. We defined the gain as the slope at the inflection point of the sigmoid.

4.7 Single-cell response and selectivity

We estimated the proportion of neurons that were significantly excited or inhibited by artificial perturbations (see 3a-b) or locomotion (see 6c) during periods of ongoing activity, in the absence of sensory stimuli. In the model, we simulated 40 trials per network, for 10 networks per each value of the perturbation. each trial in the interval $[-0.5, 1]$s, with onset of the perturbation at $t = 0$ (the perturbation was modeled as a double exponential with rise and decay times $[0.2, 1]$, Fig. 3a). In the data, we estimated spike counts after matching sample size between rest and running conditions. Spike counts were estimated in 500ms windows for each neuron in both behavioral conditions, and significant difference between the conditions was assessed with a rank-sum test.

We estimated single neuron selectivity to sensory stimuli in each condition from the average firing rate responses $r_i^a(t)$ of the i-th neuron to stimulus a in trial t. For each pair of stimuli, selectivity was estimated as

$$d'(a,b) = \frac{\text{mean}\left[r(t)^a\right] - \text{mean}\left[r(t)^b\right]}{\sqrt{\frac{1}{2}\left(\text{var}[r(t)^a] + \text{var}[r(t)^b]\right)}} \ ,$$

where mean and var are estimated across trials. The d' was then averaged across stimulus pairs.

Acknowledgments

This work was supported by a National Institute of Deafness and Other Communication Disorders Grant K25-DC013557. We would like to thank G. La Camera, Y. Ahmadian, M.

Wehr for useful discussions and suggestions.

References

[1] Pekka Niemi and Risto Näätänen. Foreperiod and simple reaction time. *Psychological bulletin*, 89(1):133, 1981.

[2] Walter E Grueninger and Karl H Pribram. Effects of spatial and nonspatial distractors on performance latency of monkeys with frontal lesions. *Journal of comparative and physiological psychology*, 68(2p1):203, 1969.

[3] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[4] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[5] Santiago Jaramillo and Anthony Zador. Auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nature Precedings*, pages 1–1, 2010.

[6] Chad L Samuelsen, Matthew PH Gardner, and Alfredo Fontanini. Effects of cue-triggered expectation on cortical processing of taste. *Neuron*, 74(2):410–422, 2012.

[7] A Moses Lee, Jennifer L Hoy, Antonello Bonci, Linda Wilbrecht, Michael P Stryker, and Cristopher M Niell. Identification of a brainstem circuit regulating visual cortical state in parallel with locomotion. *Neuron*, 83(2):455–466, 2014.

[8] Lucas Pinto, Michael J Goard, Daniel Estandian, Min Xu, Alex C Kwan, Seung-Hee Lee, Thomas C Harrison, Guoping Feng, and Yang Dan. Fast modulation of visual perception by basal forebrain cholinergic neurons. *Nature neuroscience*, 16(12):1857–1863, 2013.

[9] Yu Fu, Jason M Tucciarone, J Sebastian Espinosa, Nengyin Sheng, Daniel P Darcy, Roger A Nicoll, Z Josh Huang, and Michael P Stryker. A cortical circuit for gain control by behavioral state. *Cell*, 156(6):1139–1152, 2014.

[10] Zengcai V Guo, Nuo Li, Daniel Huber, Eran Ophir, Diego Gutnisky, Jonathan T Ting, Guoping Feng, and Karel Svoboda. Flow of cortical activity underlying a tactile decision in mice. *Neuron*, 81(1):179–194, 2014.

[11] Tsai-Wen Chen, Nuo Li, Kayvon Daie, and Karel Svoboda. A map of anticipatory activity in mouse motor cortex. *Neuron*, 94(4):866–879, 2017.

[12] Anders Nelson, David M Schneider, Jun Takatoh, Katsuyasu Sakurai, Fan Wang, and Richard Mooney. A circuit for motor cortical modulation of auditory cortical activity. *Journal of Neuroscience*, 33(36):14342–14353, 2013.

[13] Marcus Leinweber, Daniel R Ward, Jan M Sobczak, Alexander Attinger, and Georg B Keller. A sensorimotor circuit in mouse cortex for visual flow predictions. *Neuron*, 95(6):1420–1432, 2017.

[14] Siyu Zhang, Min Xu, Tsukasa Kamigaki, Johnny Phong Hoang Do, Wei-Cheng Chang, Sean Jenvay, Kazunari Miyamichi, Liqun Luo, and Yang Dan. Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, 345(6197):660–665, 2014.

[15] Matthew J McGinley, Martin Vinck, Jacob Reimer, Renata Batista-Brito, Edward Zagha, Cathryn R Cadwell, Andreas S Tolias, Jessica A Cardin, and David A McCormick. Waking

state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015.

[16] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594, 2009.

[17] Maria C Dadarlat and Michael P Stryker. Locomotion enhances neural encoding of visual stimuli in mouse v1. *Journal of Neuroscience*, 37(14):3764–3775, 2017.

[18] Jacob Reimer, Matthew J McGinley, Yang Liu, Charles Rodenkirch, Qi Wang, David A McCormick, and Andreas S Tolias. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, 7(1):1–7, 2016.

[19] Edward Zagha, Xinxin Ge, and David A McCormick. Competing neural ensembles in motor cortex gate goal-directed motor output. *Neuron*, 88(3):565–577, 2015.

[20] Luca Mazzucato, Giancarlo La Camera, and Alfredo Fontanini. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nature neuroscience*, page 1, 2019.

[21] Chengcheng Huang, Douglas A Ruff, Ryan Pyle, Robert Rosenbaum, Marlene R Cohen, and Brent Doiron. Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348, 2019.

[22] D. J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex*, 7(3):237–52, 1997.

[23] A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat Neurosci*, 15(11):1498–505, 2012.

[24] G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS Comput Biol*, 8(3):e1002395, 2012.

[25] Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Dynamics of multistable states during ongoing and evoked cortical activity. *The Journal of Neuroscience*, 35(21):8214–8231, 2015.

[26] Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience*, 10:11, 2016.

[27] Vahid Rostami, Thomas Rost, Alexa Riehle, Sacha J van Albada, and Martin P Nawrot. Spiking neural network model of motor cortex with joint excitatory and inhibitory clusters reflects task uncertainty, reaction times, and variability dynamics. *bioRxiv*, 2020.

[28] Michael T Schaub, Yazan N Billeh, Costas A Anastassiou, Christof Koch, and Mauricio Barahona. Emergence of slow-switching assemblies in structured neuronal networks. *PLoS computational biology*, 11(7), 2015.

[29] Roozbeh Kiani, Christopher J Cueva, John B Reppas, Diogo Peixoto, Stephen I Ryu, and William T Newsome. Natural grouping of neural responses reveals spatially segregated clusters in prearcuate cortex. *Neuron*, 85(6):1359–1373, 2015.

[30] Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS biology*, 3(3), 2005.

[31] Rodrigo Perin, Thomas K Berger, and Henry Markram. A synaptic organizing principle for

cortical neuronal groups. *Proceedings of the National Academy of Sciences*, 108(13):5419–5424, 2011.

[32] Wei-Chung Allen Lee, Vincent Bonin, Michael Reed, Brett J Graham, Greg Hood, Katie Glattfelder, and R Clay Reid. Anatomy and function of an excitatory network in the visual cortex. *Nature*, 532(7599):370–374, 2016.

[33] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17(11):4382–4388, 1997.

[34] M. Mascaro and D. J. Amit. Effective neural response function for collective population states. *Network*, 10(4):351–73, 1999.

[35] M. Mattia and M. V. Sanchez-Vives. Exploring the spectrum of dynamical regimes and timescales in spontaneous cortical activity. *Cogn Neurodyn*, 6(3):239–50, 2012.

[36] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.

[37] Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg, and Nicolas Brunel. Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience*, 18(12):1804, 2015.

[38] Stefano Recanatesi, Ulises Pereira, Masayoshi Murakami, Zachary Mainen, and Luca Mazzucato. Metastable attractors explain the variable timing of stable behavioral action sequences. *bioRxiv*, 2020.

[39] Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. A survey of spiking activity reveals a functional hierarchy of mouse corticothalamic visual areas. *bioRxiv*, page 805010, 2019.

[40] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

[41] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, page 1, 2019.

[42] Mario Dipoppa, Adam Ranson, Michael Krumin, Marius Pachitariu, Matteo Carandini, and Kenneth D Harris. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron*, 98(3):602–615, 2018.

[43] M. Abeles, H. Bergman, I. Gat, I. Meilijson, E. Seidemann, N. Tishby, and E. Vaadia. Cortical activity flips among quasi-stationary states. *Proc Natl Acad Sci USA*, 92:8616–8620, 1995.

[44] L. M. Jones, A. Fontanini, B. F. Sadacca, P. Miller, and D. B. Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc Natl Acad Sci U S A*, 104(47):18772–7, 2007.

[45] Tatiana A Engel, Nicholas A Steinmetz, Marc A Gieselmann, Alexander Thiele, Tirin Moore, and Kwabena Boahen. Selective modulation of cortical state during spatial attention. *Science*, 354(6316):1140–1144, 2016.

[46] A. Ponce-Alvarez, V. Nacher, R. Luna, A. Riehle, and R. Romo. Dynamics of cortical

neuronal ensembles transit from decision making to storage for later report. *J Neurosci*, 32(35):11956–69, 2012.

[47] Kourosh Maboudi, Etienne Ackermann, Laurel Watkins de Jong, Brad E Pfeiffer, David Foster, Kamran Diba, and Caleb Kemere. Uncovering temporal structure in hippocampal output patterns. *eLife*, 7:e34467, 2018.

[48] Erin L Rich and Jonathan D Wallis. Decoding subjective decisions from orbitofrontal cortex. *Nature neuroscience*, 19(7):973, 2016.

[49] Brian F Sadacca, Narendra Mukherjee, Tony Vladusich, Jennifer X Li, Donald B Katz, and Paul Miller. The behavioral relevance of cortical neural ensemble responses emerges suddenly. *Journal of Neuroscience*, 36(3):655–669, 2016.

[50] Jalil Taghia, Weidong Cai, Srikanth Ryali, John Kochalka, Jonathan Nicholas, Tianwen Chen, and Vinod Menon. Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nature communications*, 9(1):2505, 2018.

[51] Gustavo Deco, Josephine Cruzat, Joana Cabral, Enzo Tagliazucchi, Helmut Laufs, Nikos K Logothetis, and Morten L Kringelbach. Awakening: Predicting external stimulation to force transitions between different brain states. *Proceedings of the National Academy of Sciences*, 116(36):18088–18097, 2019.

[52] P. Miller and D. B. Katz. Stochastic transitions between neural states in taste processing and decision-making. *J Neurosci*, 30(7):2559–70, 2010.

[53] A. Litwin-Kumar and B. Doiron. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nat Commun*, 5:5319, 2014.

[54] M. M. Churchland, B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani, and K. V. Shenoy. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat Neurosci*, 13(3):369–78, 2010.

[55] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592, 2009.

[56] A Sanzeni, Bradley Akitake, Hannah C Goldbach, Caitlin E Leedy, Nicolas Brunel, and Mark H Histed. Inhibition stabilization is a widespread property of cortical networks. *bioRxiv*, page 656710, 2019.

[57] Alexandra K Moore, Aldis P Weible, Timothy S Balmer, Laurence O Trussell, and Michael Wehr. Rapid rebalancing of excitation and inhibition by cortical circuitry. *Neuron*, 97(6):1341–1355, 2018.

[58] M Shafi, Y Zhou, J Quintana, C Chow, J Fuster, and M Bodner. Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience*, 146(3):1082–1108, 2007.

[59] Tomáš Hromádka, Michael R DeWeese, and Anthony M Zador. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS biology*, 6(1), 2008.

[60] Daniel H O'Connor, Simon P Peron, Daniel Huber, and Karel Svoboda. Neural activity in

barrel cortex underlying vibrissa-based object localization in mice. *Neuron*, 67(6):1048–1061, 2010.

[61] Alex Roxin, Nicolas Brunel, David Hansel, Gianluigi Mongillo, and Carl van Vreeswijk. On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45):16217–16226, 2011.

[62] Jessica A Cardin, Larry A Palmer, and Diego Contreras. Cellular mechanisms underlying stimulus-dependent gain modulation in primary visual cortex neurons in vivo. *Neuron*, 59(1):150–160, 2008.

[63] Bilal Haider and David A McCormick. Rapid neocortical dynamics: cellular and network mechanisms. *Neuron*, 62(2):171–189, 2009.

[64] Carrie J McAdams and John HR Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, 19(1):431–441, 1999.

[65] Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, 1999.

[66] Neil C Rabinowitz, Robbe L Goris, Marlene Cohen, and Eero P Simoncelli. Attention stabilizes the shared gain of v4 populations. *Elife*, 4:e08998, 2015.

[67] Frances S Chance, Larry F Abbott, and Alex D Reyes. Gain modulation from background synaptic input. *Neuron*, 35(4):773–782, 2002.

[68] David A McCormick. Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Progress in neurobiology*, 39(4):337–388, 1992.

[69] Michael Goard and Yang Dan. Basal forebrain activation enhances cortical coding of natural scenes. *Nature neuroscience*, 12(11):1444, 2009.

[70] Emmanuel Eggermann, Yves Kremer, Sylvain Crochet, and Carl CH Petersen. Cholinergic signals in mouse barrel cortex during active whisker sensing. *Cell reports*, 9(5):1654–1660, 2014.

[71] Guillaume P Dugué, Magor L Lörincz, Eran Lottem, Enrica Audero, Sara Matias, Patricia A Correia, Clément Léna, and Zachary F Mainen. Optogenetic recruitment of dorsal raphe serotonergic neurons acutely decreases mechanosensory responsivity in behaving mice. *PloS one*, 9(8), 2014.

[72] Laura M Hurley, Ann M Thompson, and George D Pollak. Serotonin in the inferior colliculus. *Hearing research*, 168(1-2):1–11, 2002.

[73] Eran Lottem, Magor L Lörincz, and Zachary F Mainen. Optogenetic activation of dorsal raphe serotonin neurons rapidly inhibits spontaneous but not odor-evoked activity in olfactory cortex. *Journal of Neuroscience*, 36(1):7–18, 2016.

[74] Angie M Michaiel, Philip RL Parker, and Cristopher M Niell. A hallucinogenic serotonin-2a receptor agonist reduces visual response gain and alters temporal dynamics in mouse v1. *Cell reports*, 26(13):3475–3483, 2019.

[75] Daniel J Felleman and DC Essen Van. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[76] Leopoldo Petreanu, Diego A Gutnisky, Daniel Huber, Ning-long Xu, Dan H O'Connor, Lin Tian, Loren Looger, and Karel Svoboda. Activity in motor–sensory projections reveals distributed coding in somatosensation. *Nature*, 489(7415):299, 2012.

[77] Edward Zagha, Amanda E Casale, Robert NS Sachdev, Matthew J McGinley, and David A McCormick. Motor cortex feedback influences sensory processing by modulating network state. *Neuron*, 79(3):567–578, 2013.

[78] David B Salkoff, Edward Zagha, Erin McCarthy, and David A McCormick. Movement and performance explain widespread cortical activity in a visual detection task. *Cerebral Cortex*, 30(1):421–437, 2020.

[79] Roberto Vincis and Alfredo Fontanini. Associative learning changes cross-modal representations in the gustatory cortex. *Elife*, 5:e16420, 2016.

[80] James J Knierim and David C Van Essen. Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *Journal of neurophysiology*, 67(4):961–980, 1992.

[81] John H Reynolds and David J Heeger. The normalization model of attention. *Neuron*, 61(2):168–185, 2009.

[82] Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

[83] Benjamin R Arenkiel, Joao Peca, Ian G Davison, Catia Feliciano, Karl Deisseroth, George J Augustine, Michael D Ehlers, and Guoping Feng. In vivo light-induced activation of neural circuitry in transgenic mice expressing channelrhodopsin-2. *Neuron*, 54(2):205–218, 2007.

[84] Nuo Li, Susu Chen, Zengcai V Guo, Han Chen, Yan Huo, Hidehiko K Inagaki, Guang Chen, Courtney Davis, David Hansel, Caiying Guo, et al. Spatiotemporal constraints on optogenetic inactivation in cortical circuits. *Elife*, 8, 2019.

[85] Christopher Ebsch and Robert Rosenbaum. Imbalanced amplification: A mechanism of amplification and suppression from local imbalance of excitation and inhibition in cortical circuits. *PLoS computational biology*, 14(3):e1006048, 2018.

[86] Alexandre Mahrach, Guang Chen, Nuo Li, Carl van Vreeswijk, and David Hansel. Mechanisms underlying the response of mouse cortical networks to optogenetic manipulation. *eLife*, 9, 2020.

[87] Timothy M Otchy, Steffen BE Wolff, Juliana Y Rhee, Cengiz Pehlevan, Risa Kawai, Alexandre Kempf, Sharon MH Gobes, and Bence P Ölveczky. Acute off-target effects of neural circuit manipulations. *Nature*, 528(7582):358–363, 2015.

[88] Boaz Keren-Zur, Luca Mazzucato, and Yaron Oz. Direct mediation and a visible metastable supersymmetry breaking sector. *Journal of High Energy Physics*, 2008(10):099, 2008.

[89] Aman B Saleem, Aslı Ayaz, Kathryn J Jeffery, Kenneth D Harris, and Matteo Carandini. Integration of visual motion and locomotion in mouse visual cortex. *Nature neuroscience*, 16(12):1864–1869, 2013.

[90] Aslı Ayaz, Aman B Saleem, Marieke L Schölvinck, and Matteo Carandini. Locomotion controls spatial integration in mouse visual cortex. *Current Biology*, 23(10):890–894, 2013.

[91] Pierre-Olivier Polack, Jonathan Friedman, and Peyman Golshani. Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience*, 16(9):1331, 2013.

[92] N. Fourcaud and N. Brunel. Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Comput*, 14(9):2057–110, 2002.

[93] Ahmad Jezzini, Luca Mazzucato, Giancarlo La Camera, and Alfredo Fontanini. Processing of hedonic and chemosensory features of taste in medial prefrontal and insular networks. *The Journal of Neuroscience*, 33(48):18966–18978, 2013.

[94] Ulises Pereira and Nicolas Brunel. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 99(1):227–238, 2018.

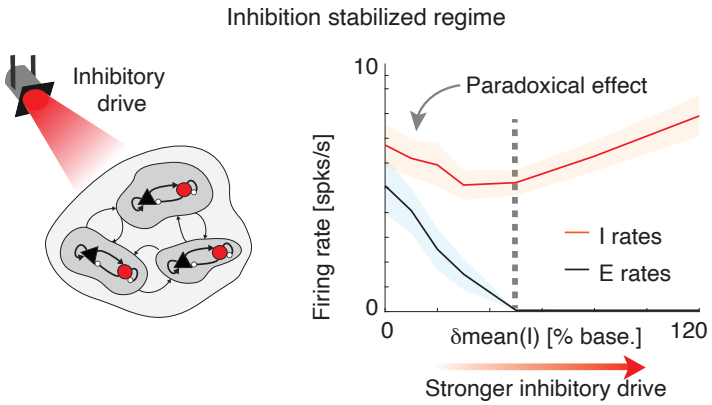

Figure S1. Inhibition stabilization in clustered networks. When increasing the inhibitory drive (afferent current to the I population), both E and I firing rates decrease (black and red curve in right panel, mean±s.e.m. across 10 simulated networks), realizing the paradoxical effect, signature of the inhibition stabilized regime [33]. Beyond δmean(I)=50% the E population shuts down and the I population rebounds (dashed vertical line).

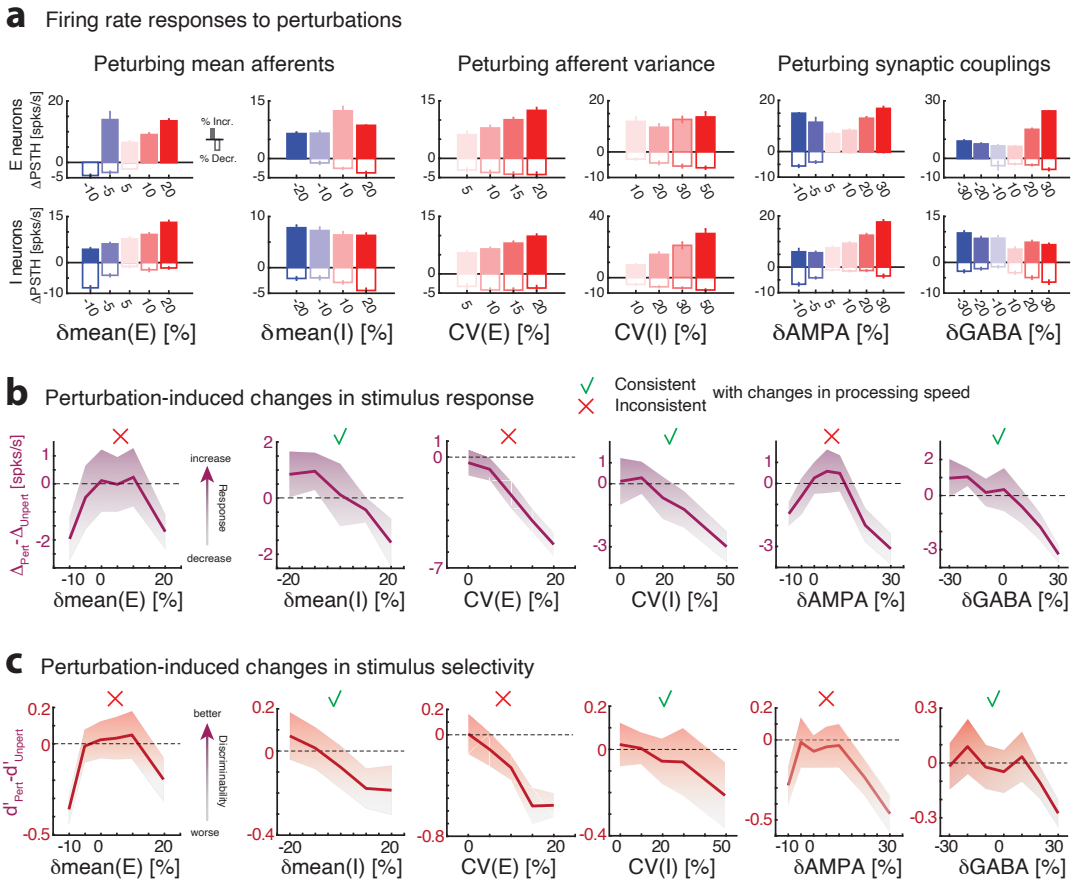

Figure S2. Perturbation-induced changes in single-cell response. **a**) Changes in peak-firing rate compared to baseline (ΔPSTH=peak-baseline; positive for excited responses, negative for inhibited responses) for E and I responsive neurons (fractions reported in Fig. 3b), in response to a perturbation with time course as in Fig. 3a. **b**) Single-cell changes in firing rate response to stimuli due to the perturbations (Δ =peak response-baseline in each perturbed or unperturbed condition) are overall uncorrelated to changes in stimulus-decoding latencies (mean±s.e.m. across 10 networks; cfr. Fig. 3c-d). **c**) Single-cell changes in stimulus selectivity due to the perturbations (d') are overall uncorrelated to changes in stimulus-decoding latencies (same notation as in **a**). While for perturbations affecting I populations (green check mark) single-cell responses are correlated to changes in coding speed, for perturbations affecting the E populations (red cross) single cell-responses are not correlated to changes in coding speed.

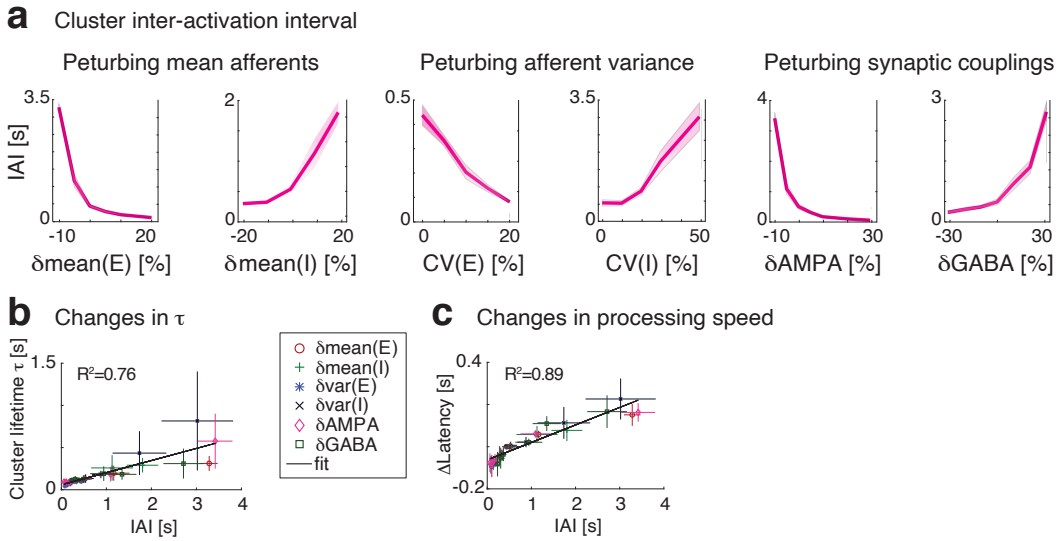

Figure S3. Perturbations-induced modulations in timescales. Perturbation-induced changes in the cluster inter-activation interval (IAI, **a**) closely track the changes in cluster activation lifetime τ (**b**) and correlate strongly with the perturbation-induced changes in stimulus-processing speed (**c**).

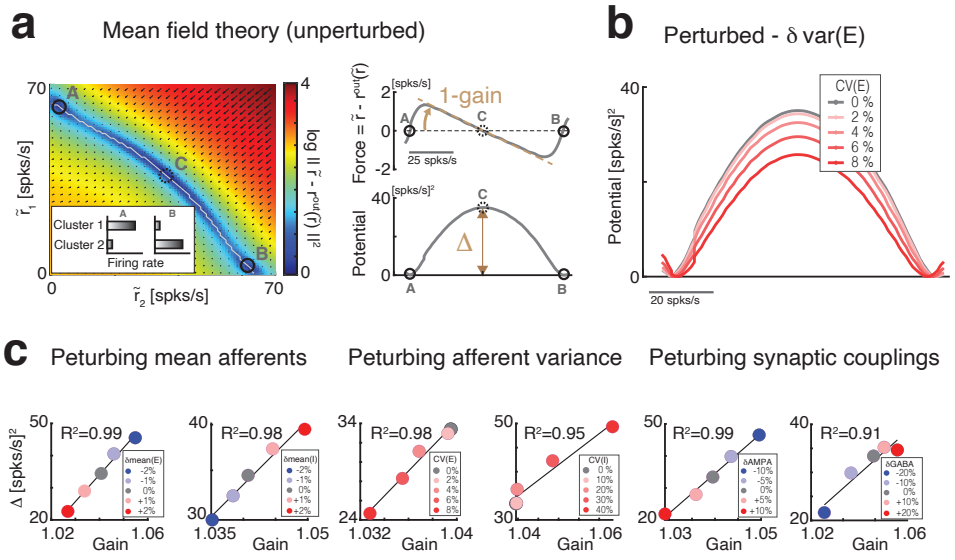

Figure S4. Effective mean field theory for barrier height and gain. **a**) Left: Reduced two-cluster theory showing the force vector (black arrows, color-coded map represents the log of the force vector norm) acting on a configuration where the two clusters have firing rates $\tilde{r}_1, \tilde{r}_2$. The force vanishes at the stable fixed points A and B, corresponding to attractors where either cluster is active and the other inactive (inset), and at saddle point C between them. Top right: From the projection of the force vector on the trajectory between the attractors (white curve in left panel) one obtains an effective transfer function $r^{out}(\tilde{r})$ whose slope yields the population intrinsic gain. Bottom right: The energy barrier separating the two attractors A and B is defined as the line integral of the projected force along the trajectory. **b**): The perturbation δvar(E) lowers the energy barrier between the two attractors (darker color-shades represent increasing values CV(E) of the perturbation). **c**) Mean field theory predicts a direct relationship between the height of the barrier Δ separating the attractors and the gain for all perturbations.

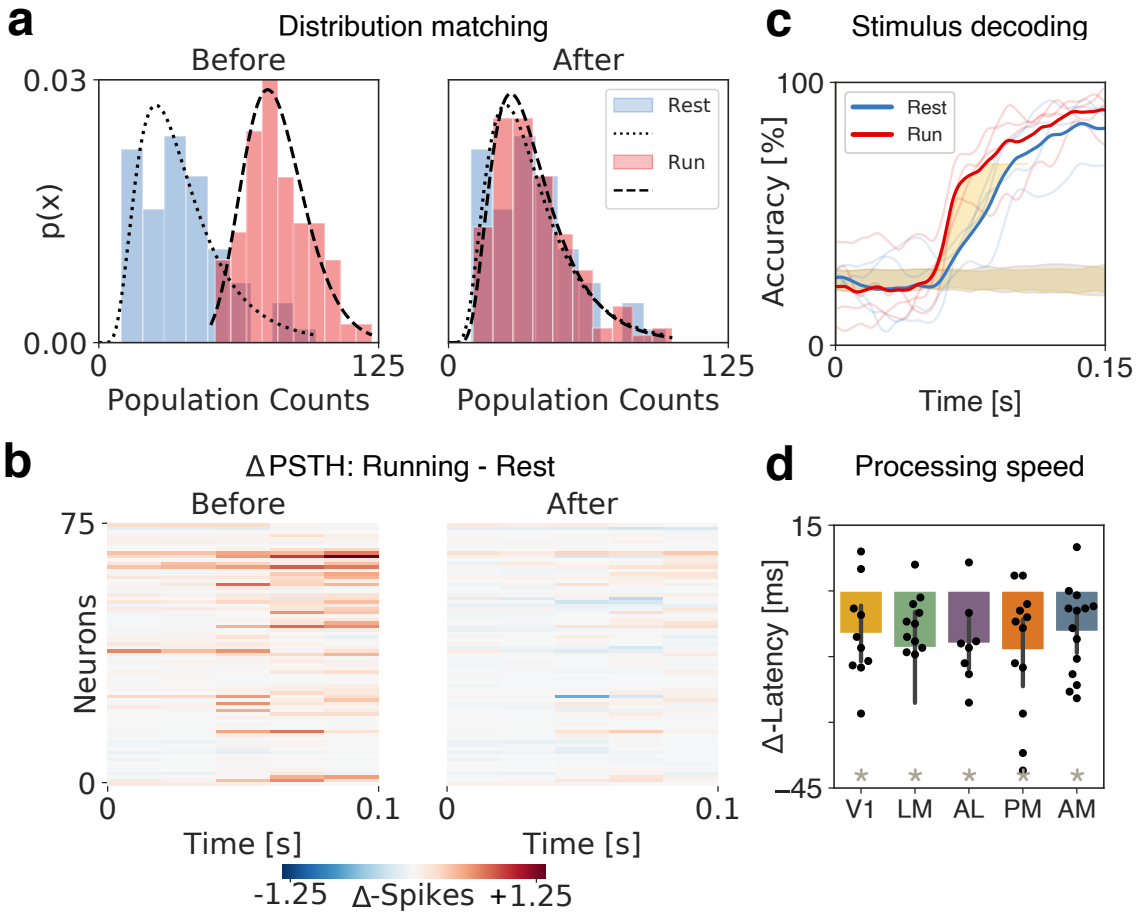

Figure S5. Anticipation of stimulus decoding persists even after matching the distribution of firing rates across behavioral conditions. **a**) Firing rate distributions for both rest and running before (left) and after (right) randomly removing spikes from the running condition. Black lines show log-normal fits of distributions. **b**) Δ PSTH between behavioral conditions before and after distribution matching shows effects of match across each neuron's firing rate. **c**) Mean stimulus-decoding accuracy across orientations per behavioral condition using neurons from V1 as predictors shows the anticipation of the stimulus in the running condition after distribution matching (same sessions as in Fig. 6e). **d**) Summary of changes in processing speed due to locomotion by area after distribution matching. (t-test, $p < 0.01$)

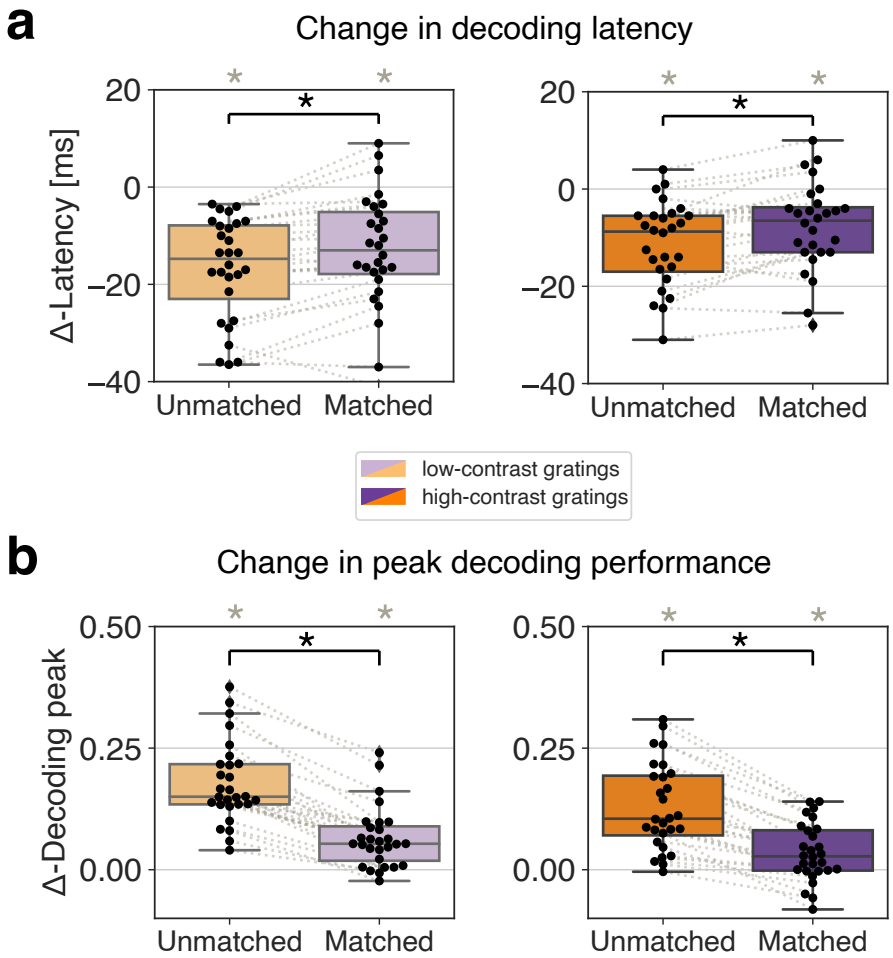

Figure S6. Matching the distribution of firing rates between behavioral conditions reduces the change in peak decoding, but preserves the change in decoding latency between behavioral conditions. **a**) Δ-Latency over all areas, separated by the grating contrast shows that even after matching the distribution of firing rates between conditions (purple), the increase in sensory processing during running was still significant (rank-sum test, gray $* = p < 0.005$) The change in Δ-Latency between non-matched (orange) and matched (purple) datasets was significant (rank-sum test,black $* = p < 0.001$) **b**) The difference in peak decoding between behavioral conditions is reduced for low and high contrast drifting grating trials after matching the distributions (rank-sum test, gray $* = p < 0.005$). The change in Δ-Decoding peak between non-matched and matched datasets was significant (rank-sum test, black $* = p < 0.005$) for both contrasts.

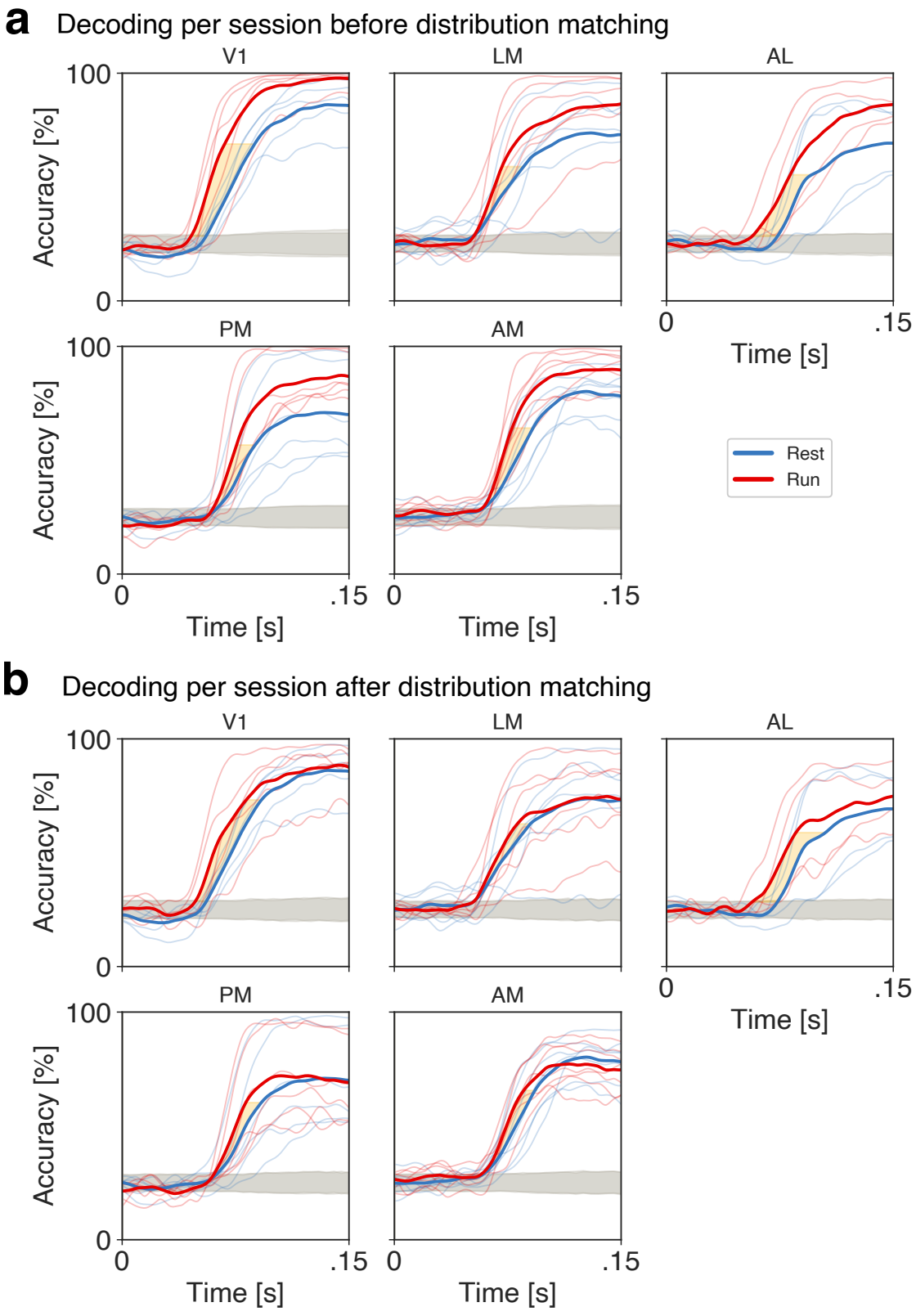

Figure S7. Mean stimulus-decoding accuracy of high-contrast drifting gratings across sessions per behavioral condition and area before (**a**) and after (**b**) matching the distribution of firing rates shows the decrease in Δ-Decoding peaks and preservation of Δ-Latency. Notations as in Fig. 6e.

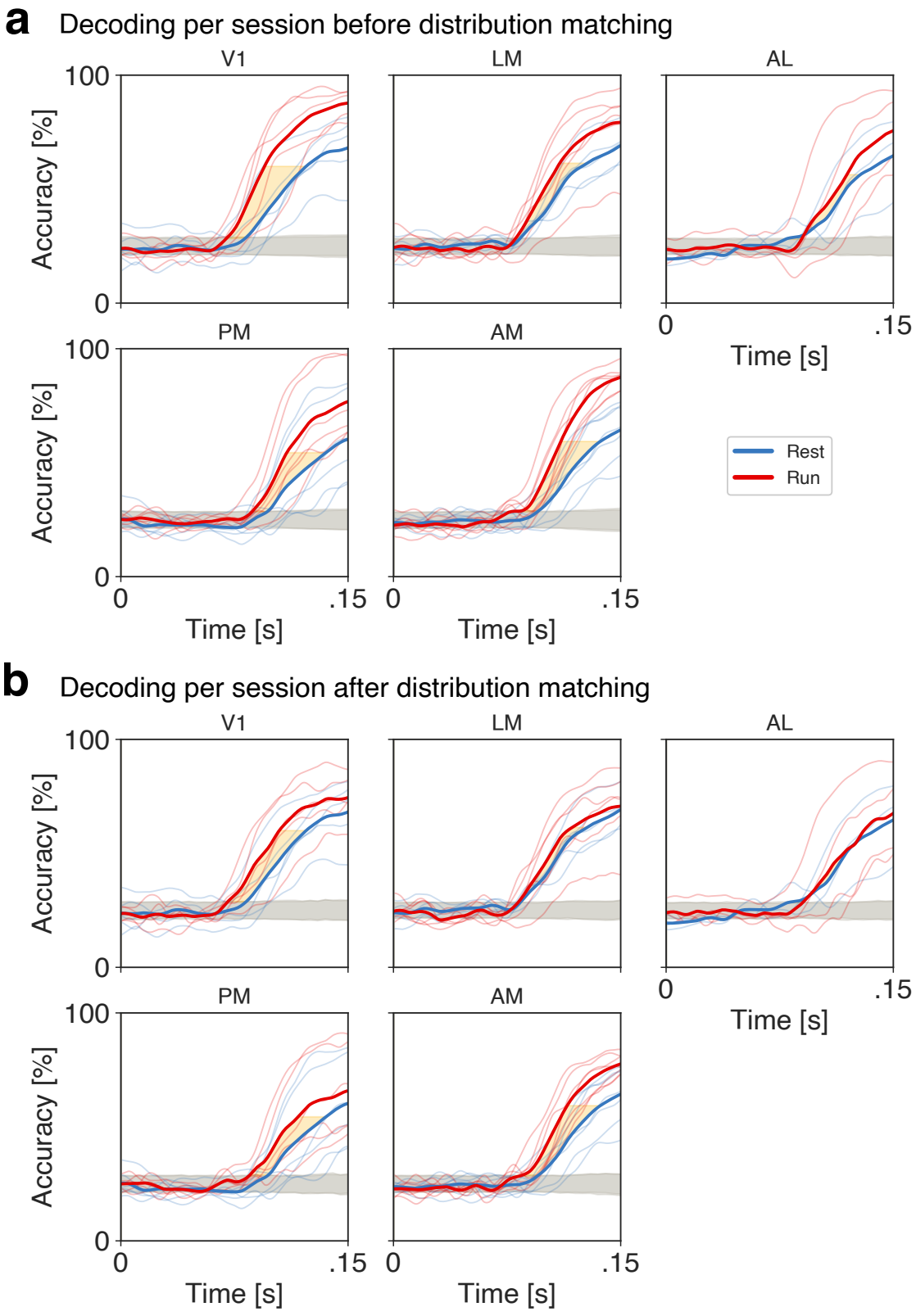

Figure S8. Mean stimulus-decoding accuracy of low-contrast drifting gratings across sessions per behavioral condition and area before (**a**) and after (**b**) matching the distribution of firing rates shows the decrease in Δ-Decoding peaks and preservation of Δ-Latency. Notations as in Fig. 6e.

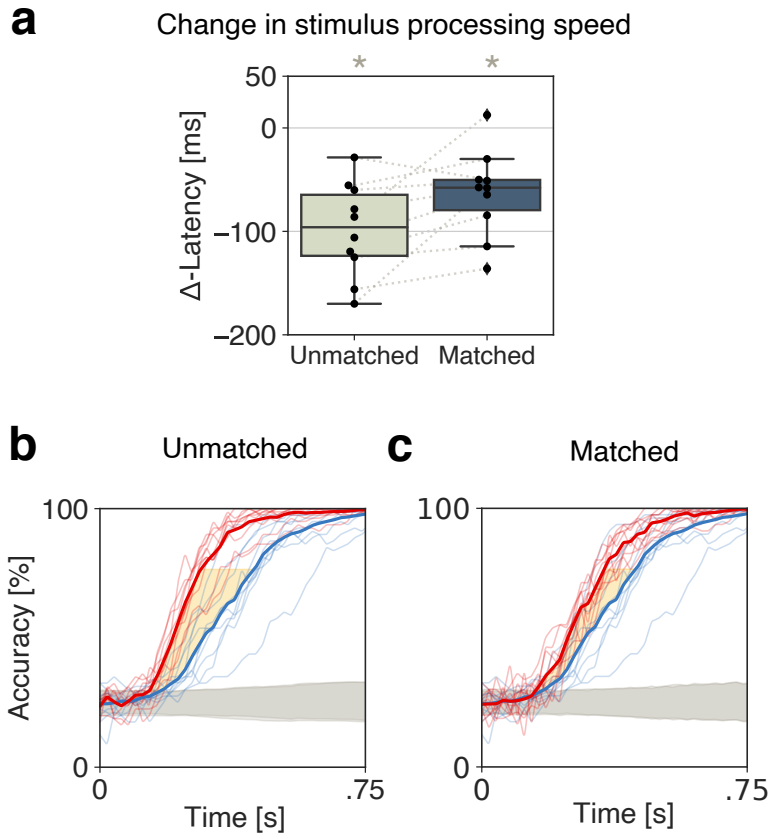

Figure S9. In the model, matching the distribution of firing rates between perturbed (δvar(E) with CV(E)=20%) and unperturbed conditions preserved the perturbation-induced acceleration in stimulus processing speed (same data as in Fig. 2b). **a**) Δ-Latency over 10 simulated networks shows that even after matching the distribution of firing rates between conditions (purple), the increase in sensory processing speed during the perturbed condition was still significant (rank-sum test, $* = p < 0.005$). There was no significant change in Δ-Latency between unmatched and matched datasets (rank-sum test,$p > 0.05$). Time course of stimulus-decoding accuracy over all 10 simulated networks before (**b**) and after (**c**) matching the distribution of firing rates.