Modulation of metastable ensemble dynamics explains the inverted-U relationship between tone discriminability and arousal in auditory cortex

Lia Papadopoulos,[1] Suhyun Jo,[1] Kevin Zumwalt,[1] Michael Wehr,[1,2]
Santiago Jaramillo,[1,3] David A. McCormick,[1,3] and Luca Mazzucato[1,3,4,5,*]

[1] *Institute of Neuroscience, University of Oregon, Eugene, OR 97403, USA*
[2] *Department of Psychology, University of Oregon, Eugene, OR 97403, USA*
[3] *Department of Biology, University of Oregon, Eugene, OR 97403, USA*
[4] *Department of Mathematics, University of Oregon, Eugene, OR 97403, USA*
[5] *Department of Physics, University of Oregon, Eugene, OR 97403, USA*

SUMMARY

Past work has reported inverted-U relationships between arousal and auditory task performance, but the underlying neural network mechanisms remain unclear. To make progress, we recorded auditory cortex activity from behaving mice during passive tone presentation and simultaneously monitored pupil-indexed arousal. In these experiments, neural discriminability of tones was maximized at intermediate arousal, revealing a neural correlate of the inverted-U. We explained this arousal-dependent sound processing using a spiking model with clusters. In the model, stimulus discriminability peaked as the network transitioned from a multi-attractor phase exhibiting slow switching between metastable cluster activations (low arousal) to a single-attractor phase with uniform activity (high arousal). This transition also qualitatively captured arousal-induced reductions of neural variability observed in the data. Altogether, this study elucidates computational principles to explain interactions between arousal, neural discriminability, and variability, and suggests that transitions in the dynamical regime of cortical networks could underlie nonlinear modulations of sensory processing.

* Lead Contact; Correspondence: lmazzuca@uoregon.edu

INTRODUCTION

Variations in brain state – such as levels of arousal – significantly impact sensory responses and information processing [1–8]. During wakefulness, increases in arousal and arousal-related neuromodulator activity are associated with increases in pupil diameter [9–11], enabling non-invasive monitoring of arousal state in behaving animals. Using pupillometry, several studies have reported an intriguing "inverted-U" relationship between baseline pupil diameter and task performance, indicating optimal performance at intermediate arousal [12–18].

The inverted-U relationship between pupil-indexed arousal and performance has been demonstrated particularly clearly in the context of auditory processing, with examples observed during sound detection and discrimination tasks in mice [12, 13] and humans [14–16]. Although some work has revealed neural substrates of optimal sound detection [12], network-level dynamical principles mediating the inverted-U relationship, especially for sound discrimination, remain unclear. Here, we combined electrophysiology and circuit modeling to shed light on potential network mechanisms underlying optimal, arousal-dependent performance states for auditory discrimination.

Given that neural signatures of inverted-U arousal-performance relationships have been observed in auditory cortex (ACtx) without task engagement [12], we examined how arousal impacted neural discriminability of pure tones during passive stimulus presentation. To this end, we used Neuropixels probes to record sound-evoked activity from ACtx of behaving mice, while simultaneously monitoring arousal levels with pupillometry. We found that tone frequency was most accurately decoded from ACtx activity during intermediate pupil dilation, in line with an inverted-U relationship. This finding extends previous results on neural correlates of optimal sound detection in ACtx [1] to population-based neural discriminability of auditory stimuli.

To illuminate potential network mechanisms governing this neural manifestation of the inverted-U relationship, we modeled ACtx as a recurrently-connected network of excitatory and inhibitory spiking neurons. We compared two canonical models for the network architecture – uniform or clustered. In the latter model, neurons were organized into strongly-connected modules representing functional neural assemblies [19]. Unlike uniformly-connected networks, clustered networks exhibit metastable activity patterns, characterized by spontaneous, transient activations of different neural assemblies [20–27].

We hypothesized that non-monotonic variations of stimulus discriminability might emerge from modulations of the metastable assembly dynamics present in clustered networks. To investigate this, we presented model networks with sensory stimulation and an arousal modulation. Motivated by experimental studies, an increase in arousal was modeled as a suppression of recurrent excitatory synaptic transmission [28–38] and an increase in external drive (representing increased thalamic activation) [12, 39–45]. Under this implementation, the clustered model reproduced the inverted-U relationship between arousal and stimulus discriminability, while the uniform model failed to display such an effect.

In the clustered model, we show that the inverted-U relationship emerges via a transition from a multi-attractor phase, characterized by slow switching between different highly-active neural assemblies (low arousal), to a single-attractor phase with uniform activity (high arousal). Optimal stimulus discriminability occurred between these two regimes, where assembly activation dynamics were present but flexibly modulated. The clustered model additionally predicted a reduction of neural variability with increasing arousal, and we found evidence of that trend in the experimental data. As a whole, our results suggest that arousal-induced transitions in the collective dynamical regime of a cortical circuit may explain certain aspects of arousal-dependent stimulus processing and neural variability in auditory cortex. Although our study does not rule out alternative explanations for these phenomena, it provides insight into one possible computational mechanism that can be further tested and built upon in future work.

RESULTS

We used Neuropixels probes to record from populations of auditory cortical neurons in head-fixed mice (primarily targeting A1), while simultaneously monitoring locomotion speed and pupil diameter (Fig. 1A; Methods). The pupil diameter in each session was normalized by its maximum value and was used as a proxy for arousal level (Fig. 1B,C). A full spectrum of arousal states was expressed in several sessions, and the middle-to-high pupil range was expressed in the remaining recordings. Single-unit activity was measured during passive sound presentation (Fig. 1D, "evoked" periods) and in the absence of auditory stimuli (Fig. 1E, "spontaneous" periods). During evoked periods, mice were presented with brief pure tones (25 ms, 2-32 kHz).

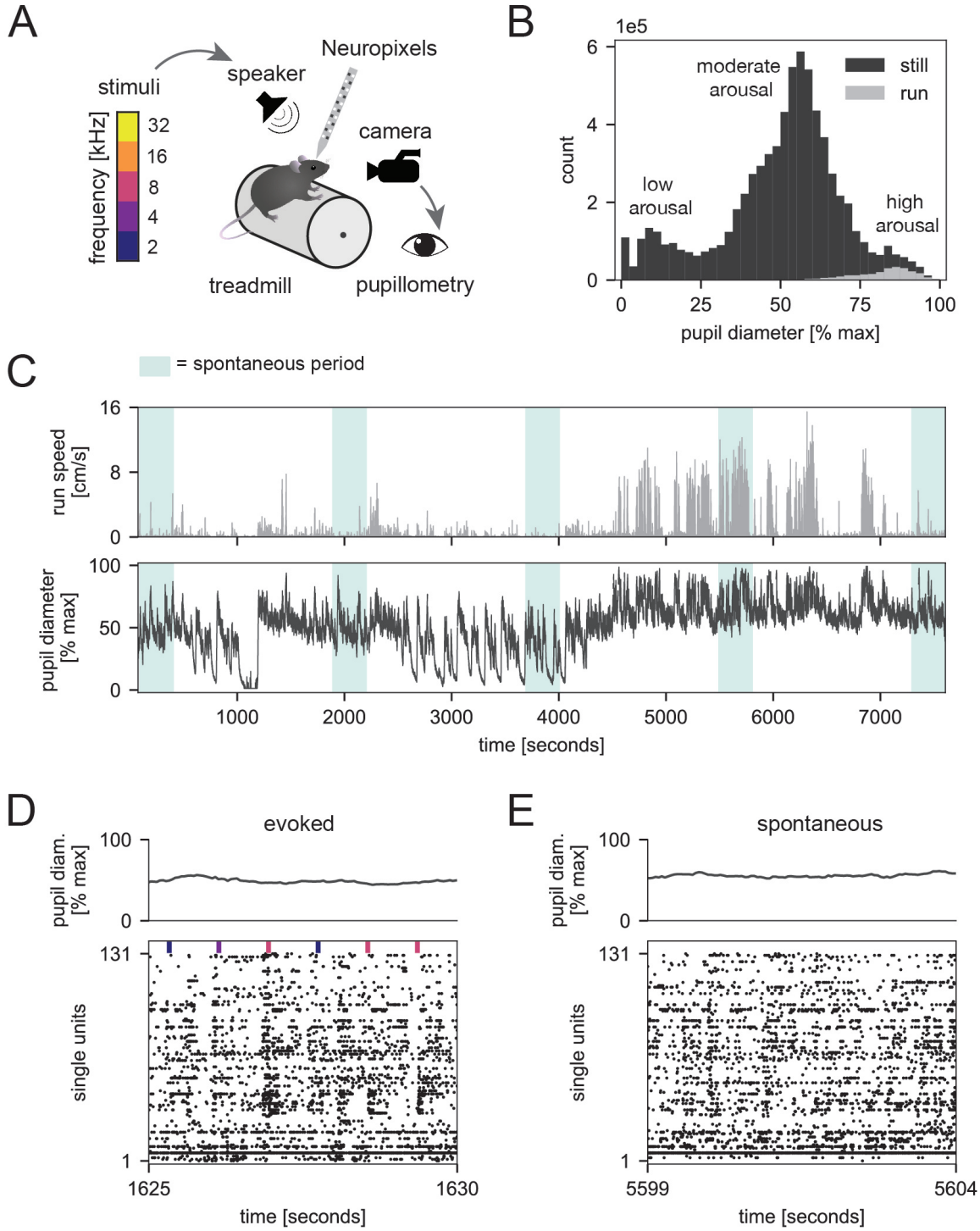

FIG. 1. **Neuropixels recordings from mouse ACtx during a range of arousal states. (A)** Schematic of experimental setup (Methods). **(B)** Pupil diameter distributions from an example session during stillness (dark gray) or running (light gray). **(C)** Running speed and pupil diameter traces from an example session. **(D)** Pupil diameter trace and population raster across 5 seconds of evoked activity; vertical ticks indicate stimulus onset times. **(E)** Same as **(D)** but for spontaneous activity.

Neural discriminability of pure tones is optimal at intermediate arousal in ACtx

To test whether arousal impacts neural discriminability of pure tones in ACtx, we split the trials in each session by pupil diameter (see Fig. 2A for an example). Within each pupil-based partition, we then computed a single-cell

measure of neural stimulus discriminability (D'_{sc}; Fig. 2B,C, Methods). On average across sessions, the cell-averaged D'_{sc} followed an inverted-U relationship with normalized pupil diameter (Fig. 2F). At the population level, intermediate pupil diameters were associated with significant increases in D'_{sc} relative to small or large diameters (Fig. 2G), and in individual sessions, the cell-averaged D'_{sc} was highest at moderate pupil diameters (Fig. S1A).

Auditory information is also encoded in the collective activity of neuronal ensembles [46–52]. To understand how arousal might affect the ability of downstream areas to read out sound information from ACtx ensemble activity, we trained an ideal-observer linear classifier to decode tone frequency from single-trial population spike-counts (Fig. 2D, Methods). Training and testing was done separately for each pupil-based partition in a session, and results were summarized with the cross-validated decoding accuracy (Fig. 2E).

On average across sessions, decoding performance exhibited an inverted-U relationship with pupil diameter

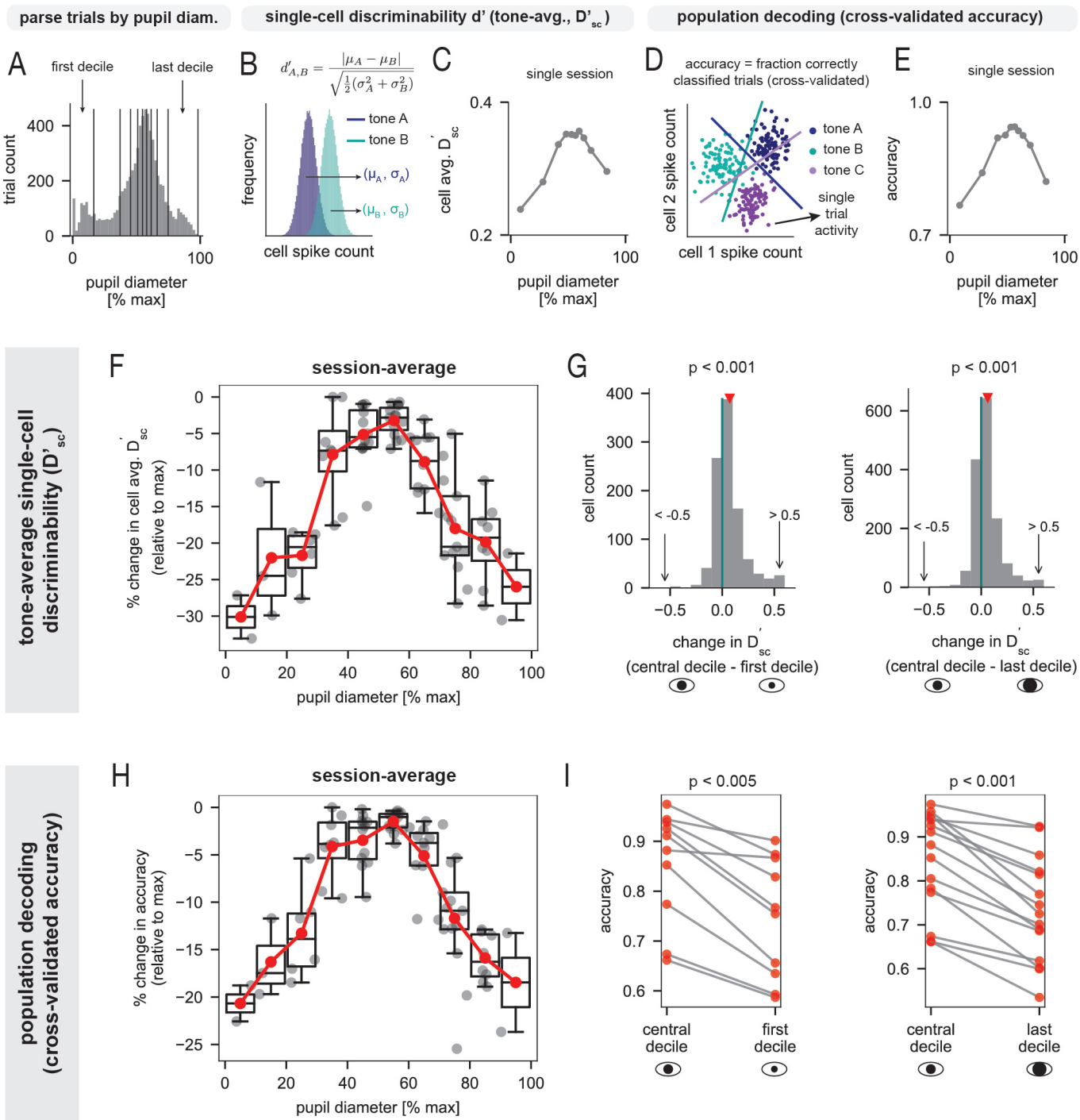

FIG. 2. **Neural discriminability of pure tones is optimal at intermediate arousal in ACtx.** **(A)** Histogram of pre-stimulus pupil diameter from an example session; black lines indicate deciles. **(B)** Schematic of single-cell discriminability index (d') for one pair of stimuli. An overall measure, D'_{sc}, was computed by averaging d' across all tone pairs (Methods). **(C)** Cell-averaged D'_{sc} in each pupil decile from **(A)**. **(D)** Schematic of linear population decoding analysis (Methods). **(E)** Cross-validated decoding accuracy in each pupil decile from **(A)**. **(F)** Percent change in cell-averaged D'_{sc} (relative to session-maximum) *vs.* pupil diameter (group data from $n = 15$ sessions). In each pupil diameter bin, we show single-session data (gray) with the corresponding boxplot and session-average (red). **(G)** *Left:* Difference in D'_{sc} between the most central and first pupil decile of a session ($n = 1002$ units from 10 sessions with average pupil diameter of first decile $\leq 33\%$ max dilation; red triangle indicates mean difference; $p < 0.001$, Wilcoxon signed-rank test). *Right:* Distribution of the difference in D'_{sc} between the most central and last pupil decile of a session ($n = 1552$ units from 15 sessions with average pupil diameter of last decile $\geq 67\%$ max dilation; red triangle indicates mean difference; $p < 0.001$, Wilcoxon signed-rank test). **(H)** Same as **(F)** but for cross-validated decoding accuracy. **(I)** *Left:* Accuracy in the most central and first pupil decile of a session (data from $n = 10$ sessions with average pupil diameter of first decile $\leq 33\%$ max dilation; $p < 0.005$, Wilcoxon signed-rank test). *Right:* Accuracy in the most central and last pupil decile of a session (data from $n = 15$ sessions with average pupil diameter of last decile $\geq 67\%$ max dilation; $p < 0.001$, Wilcoxon signed-rank test). See also Fig. S1.

(Fig. 2H), and there was a statistically significant increase in accuracy in states of moderate pupil dilation relative to more constricted or highly-dilated pupil states (Fig. 2I). Moreover, the best performance in all sessions was achieved at intermediate pupil diameters, and the worst performance at relatively small or large diameters (Fig. S1B). The session-averaged decoding performance still followed an inverted-U after excluding locomotion trials (Fig. S1C,D), though the right-hand-side decline was less pronounced. This may be due to the fact that average pupil diameters were smaller without movement data (Fig. S1E). The decoding results were also similar when pupil diameter was normalized by the global maximum across all sessions (Fig. S1F,G). Altogether, our findings reveal neural signatures of an inverted-U relationship between arousal level and neural sound discriminability at single-cell and population levels.

Network modeling of arousal-dependent stimulus processing in ACtx

What circuit mechanisms can explain the inverted-U relationship between tone discriminability and arousal in ACtx? Because the inverted-U relationship is nonlinear, we speculated that it could stem from a modulation of collective dynamics akin to a phase transition. To investigate this, we modeled ACtx as a recurrently-connected network of excitatory (E) and inhibitory (I) integrate-and-fire neurons (Fig. 3A,B; Methods). This type of spiking model strikes a balance between biological plausibility and tractability, and also allows for analysis of spiking variability (Fig. 8).

The collective activity of a cortical network depends on the architecture of synaptic couplings. Here, we considered a model with "clustered" architecture (Fig. 3A) and compared it to a control with "uniform" architecture (Fig. 3B). By comparing these alternative models, we aimed to elucidate potential dynamical principles underlying the experimental observations. Following previous work, neurons in the clustered model were arranged into clusters with strong internal synaptic coupling and relatively weak coupling to other clusters [19–22, 25]. Neurons in the uniform model were instead connected randomly with homogeneous coupling strengths. In both models, auditory stimuli were implemented as external excitatory inputs (Fig. 3A,B). In the clustered model, stimulus input targeted cells in a randomly-chosen subset of the clusters, and in the uniform model, stimulus input targeted a random subset of all excitatory cells. To match the experiments, we modeled five stimuli and allowed for overlap in the cell subgroups targeted by different stimuli, in line with the fact that auditory neurons can respond to multiple tones.

As shown in past studies, the clustered model exhibits metastable attractor dynamics [19–22, 24–27]. These dynamics occur in a regime of strong intracluster coupling, where the network's state space contains a multiplicity of attractors corresponding to different configurations of active clusters (Fig. S5A,B) [21, 22, 26]. In such a regime, random fluctuations can cause transitions between attractors; this generates metastable dynamics wherein individual clusters switch between high and low firing modes over time, and collective network activity moves between different collective states characterized by different groups of simultaneously active clusters (Fig. 3A Right). Throughout the text, we use the term "multistability" to denote the property of the state space having multiple attractors, and the term "metastable" to denote the itinerant dynamics describing spontaneous transitions between attractors. We hypothesized that modulations of metastable cluster dynamics could provide a mechanism for the inverted-U relationship. In contrast to the clustered model, the uniform model has a single attractor with asynchronous activity (Fig. 3B Right).

The second aspect of the model is the arousal implementation. Experimental studies indicate that arousal modulates sensory processing through various neuromodulatory systems [in particular, via the actions of acetylcholine (Ach) and norepinephrine (NE)] and thalamocortical mechanisms [1–3, 53–60]. Given that these pathways can induce a variety of

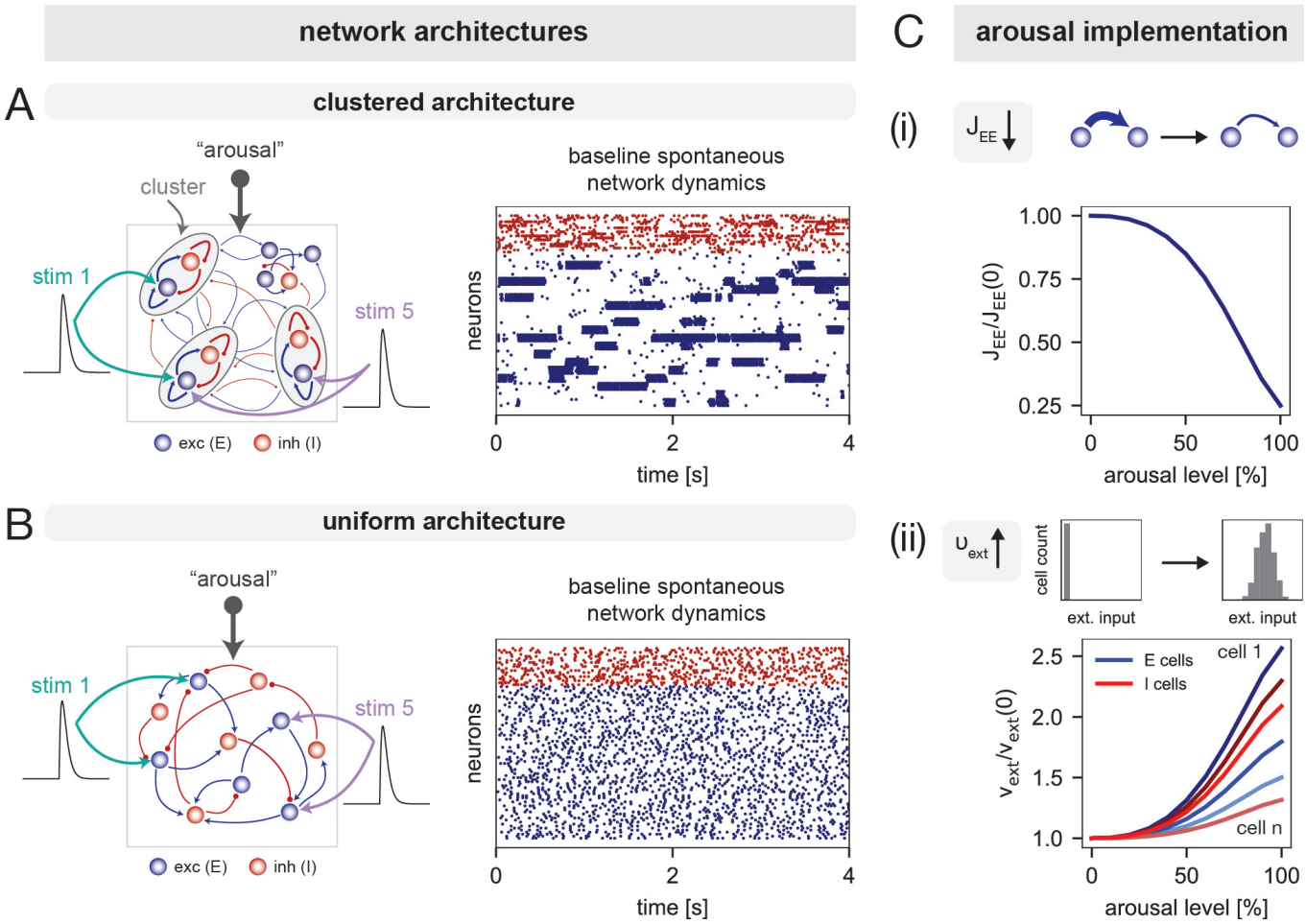

FIG. 3. **Network modeling of arousal-dependent stimulus processing in ACtx.** **(A, B)** ACtx was modeled as a recurrent network of spiking neurons arranged in either a clustered (**A Left**) or uniform (**B Left**) architecture. Both networks were subjected to an arousal signal and sensory stimulation. Raster plots show spontaneous baseline activity of a subset of neurons from a clustered (**A Right**) or uniform (**B Right**) network. **(C)** An increase in arousal was modeled as a simultaneous modulation of two parameters. **(C-i)** Synaptic strength J_{EE} (relative to baseline value) *vs.* arousal. **(C-ii)** External input ν_{ext} (relative to baseline value) *vs.* arousal; different curves correspond to different example cells. See also Table S1 and Figs. S2, S3. See Methods for model details.

effects on cortical circuits [3, 54, 55, 58–66], there are several possibilities for how to model the impact of arousal. Here, we considered one implementation involving a simultaneous modulation of two parameters. In particular, an increase in arousal was modeled as (i) a global decrease in the strength of recurrent E-to-E synapses, and (ii) an increase in the external excitatory drive to E and I cells; to improve biological plausibility, we also introduced cell-to-cell variability in the external drive modulation (Fig. 3C).

The first component of the arousal model is motivated by experiments showing that increases in Ach and NE can have suppressive effects on intracortical excitatory synaptic transmission [28–38, 67]. We reasoned that such a modulation of synaptic efficacy would strongly impact assembly dynamics in the clustered model – which rely on recurrent excitation – and thus impact stimulus processing. The second component of the arousal model is motivated by studies showing that active states are associated with increases in thalamic activity, which provides a major source of excitatory drive to sensory cortices [12, 39–45]. Importantly, this arousal implementation qualitatively captured the heterogeneity of arousal-related firing rate modulations observed in the data (Fig. S2).

In what follows, we examine how collective network dynamics are affected by the above arousal modulation. We refer to the induced changes in network dynamics as the "computational" or "network mechanism", in contrast to the "biological implementation" of arousal, which refers to the specific neurophysiological pathways and modulations of network parameters used to model arousal. Though here we examined one biological implementation of arousal supported by empirical evidence, alternative implementations can produce similar network mechanisms (see Fig. S3 and Discussion).

Evidence of functionally-organized correlation-based clusters in ACtx

A distinctive feature of the clustered model is that noise correlations – covariations in activity across repeated presentations of the same stimulus [68] – are larger between cells in the same *versus* different clusters (Fig. 4B). To better motivate the clustered model, we thus tested for the presence of functional clusters in the data. To identify putative clusters, we estimated pairwise noise correlations and performed hierarchical clustering on the resulting correlation matrix. The statistical significance of detected clusters was assessed via comparison against a trial-shuffled null model that contained no true clustering (Methods).

In the clustered model, significant functional assemblies were clearly detected with hierarchical clustering (Fig. 4A Left), and neurons were accurately assigned to their ground-truth clusters (Fig. 4B). By contrast, significant clusters were rarely identified in noise-correlation matrices from the uniform model (Fig. 4A Right). Correlations in the experimental data were typically weaker and more diffuse relative to the clustered model (Fig. 4F for two examples). Nonetheless, the clustering analysis revealed significant correlation-based assemblies (Fig. 4E). In most sessions, a majority of neurons belonged to significant clusters (Fig. 4I), and cluster sizes ranged from a few to tens of cells

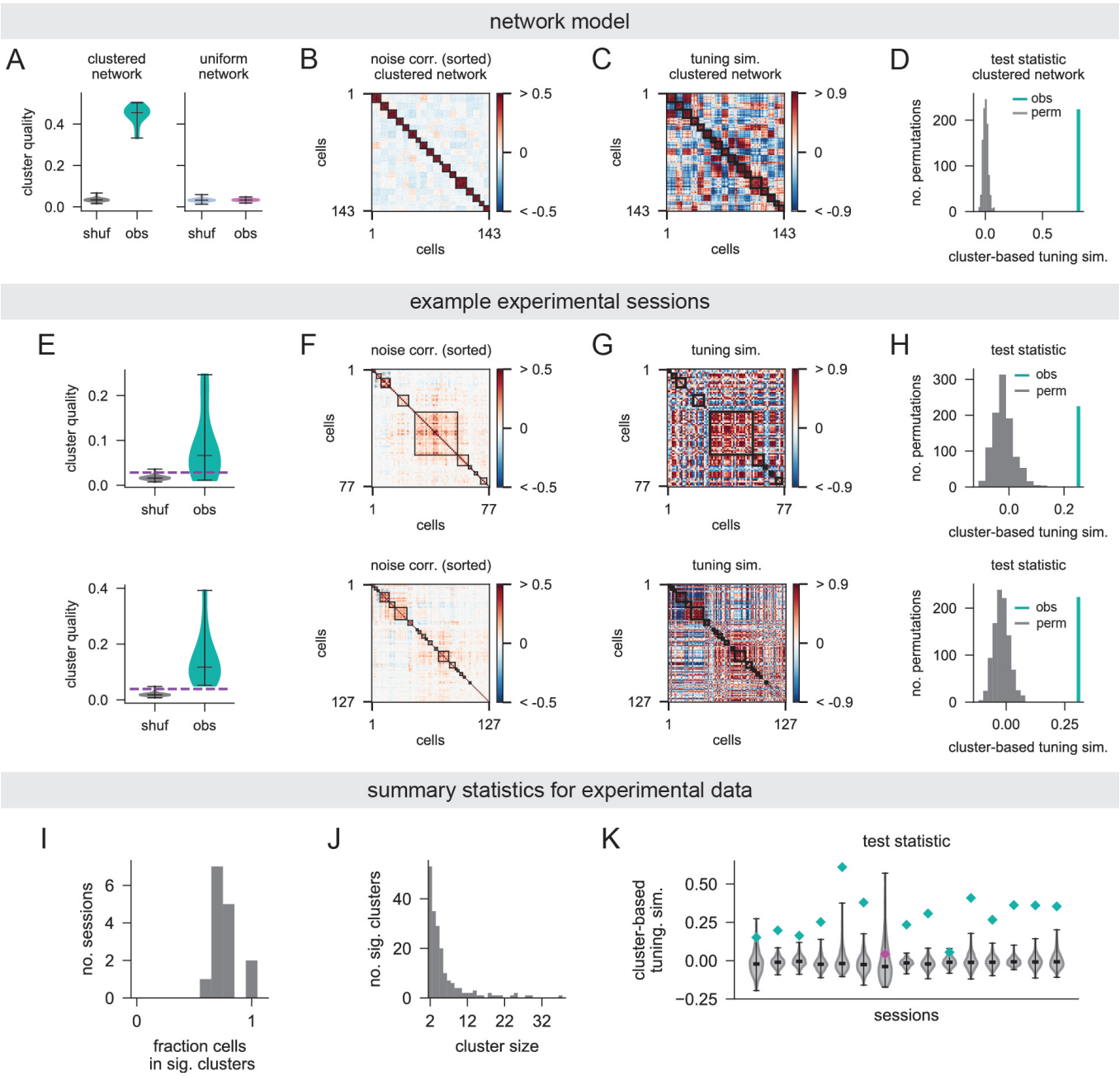

FIG. 4. **Evidence of functionally-organized correlation-based clusters in ACtx. (A-D)** Clustering analysis in the network models. **(A)** Violin plots of cluster quality obtained from hierarchical clustering of observed ("obs") and trial-shuffled ("shuf") noise-correlation matrices. In the example clustered network, all observed clusters are significant relative to trial-shuffled data ($p < 0.05$, Bonferroni-corrected); in the example uniform network, no clusters are significant. **(B)** Noise correlation matrix of a random subsample of cells from a clustered network, sorted according to detected clusters. **(C)** Tuning similarity matrix for the same cells and sorting in **(B)**. **(D)** Test-statistic for cluster-based tuning similarity, computed from the data in **(B,C)**. The observed value exceeds the distribution obtained by permuting cluster labels across cells ($p < 0.001$). **(E-H)** Clustering analysis in two experimental sessions (1 session/row). **(E)** Same as **(A)**, but for the experimental sessions. Clusters in the observed distribution above the dashed line are significant relative to trial-shuffled data ($p < 0.05$, Bonferroni-corrected). **(F)** Noise correlation matrices sorted according to detected clusters; significant clusters are outlined in black. **(G)** Tuning similarity matrices for the same cells and sorting in **(F)**. **(H)** Same as **(D)**, but for the experimental sessions. The observed values exceed the distributions obtained by permuting cluster labels across cells ($p < 0.001$). **(I-K)** Summary statistics for experimental data. **(I)** Distribution of the fraction of cells in a session that belong to significant clusters. **(J)** Distribution of cluster sizes (significant clusters only). **(K)** Test statistic for cluster-based tuning similarity in each session. Colored diamonds indicate observed values and violin plots show distributions obtained by permuting cluster labels across cells. Green indicates a significant result relative to permuted data ($p < 0.05$) and magenta otherwise. See Methods for details.

(Fig. 4J).

To assess the functional relevance of inferred clusters, we tested whether cells in the same cluster had higher-than-chance tuning similarity (Pearson correlation between trial-averaged stimulus responses [68]). For each cell, we computed the difference between its average within- and between-cluster tuning similarity, and defined a test statistic ("cluster-based tuning similarity") as the average across cells in significant clusters. The observed value of the test statistic was then compared to a null distribution obtained by randomly permuting cluster labels across neurons (Methods).

In the clustered model, cluster-based tuning similarity was significantly greater than chance (Figs. 4C,D). Because neurons in the same cluster exhibit coordinated dynamics, the presence (absence) of stimulus-related drive biases activation (inactivation) of the entire assembly, leading to similar stimulus responses for its constituent neurons. The relationship between correlation-based clusters and tuning similarity was less straightforward in the data. However, we still observed a certain degree of overlap, which was visually apparent in some sessions (Fig. 4G) and verified with the statistical test described above (Fig. 4H). In total, cluster-based tuning similarity was significant in nearly all sessions (Fig. 4K), suggesting that the detected correlation-based clusters exhibit some functional organization. These results provide additional motivation for the clustered model.

The clustered model captures the inverted-U relationship between stimulus discriminability and arousal

We next examined whether the inverted-U relationships between stimulus discriminability and arousal (Fig. 2) could be reproduced in either the uniform or clustered circuit models. We began by investigating how the single-cell discriminability index (D'_{sc}) varied with the arousal modulation (Methods). In the clustered model, the cell-averaged D'_{sc} peaked at moderate arousal (Fig. 5C), whereas in the uniform model, it decreased with arousal (Fig. 5D). Thus, at the level of single-cell discriminability, only the clustered model captured the inverted-U relationship with arousal observed in the data.

We also examined how the arousal modulation impacted the read-out of stimulus identity from population activity. To this end, we trained ideal-observer linear decoders to discriminate stimuli given responses from an ensemble of neurons sampled from the full network (Methods). Similar to single-cell discriminability, cross-validated decoding performance followed an inverted-U relationship in the clustered networks, but decreased with arousal in the uniform networks (Fig. 5E,F). Though decoding performance at moderate and high arousal did increase with population size in the clustered networks, the overall inverted-U shape of the decoding curve was relatively robust to variations in the number of sampled neurons (Fig. S4).

In sum, the clustered network can explain the inverted-U relationships between stimulus discriminability and arousal observed in the data, while the uniform network fails to do. In what follows, we examine the network mechanisms underlying the inverted-U behavior in the clustered model.

The arousal modulation controls the dynamical regime of the clustered model

Because the inverted-U relationship emerged only in the clustered networks, it must rely on an arousal-induced modulation of the ongoing metastable dynamics unique to that model. To understand the mechanism through which

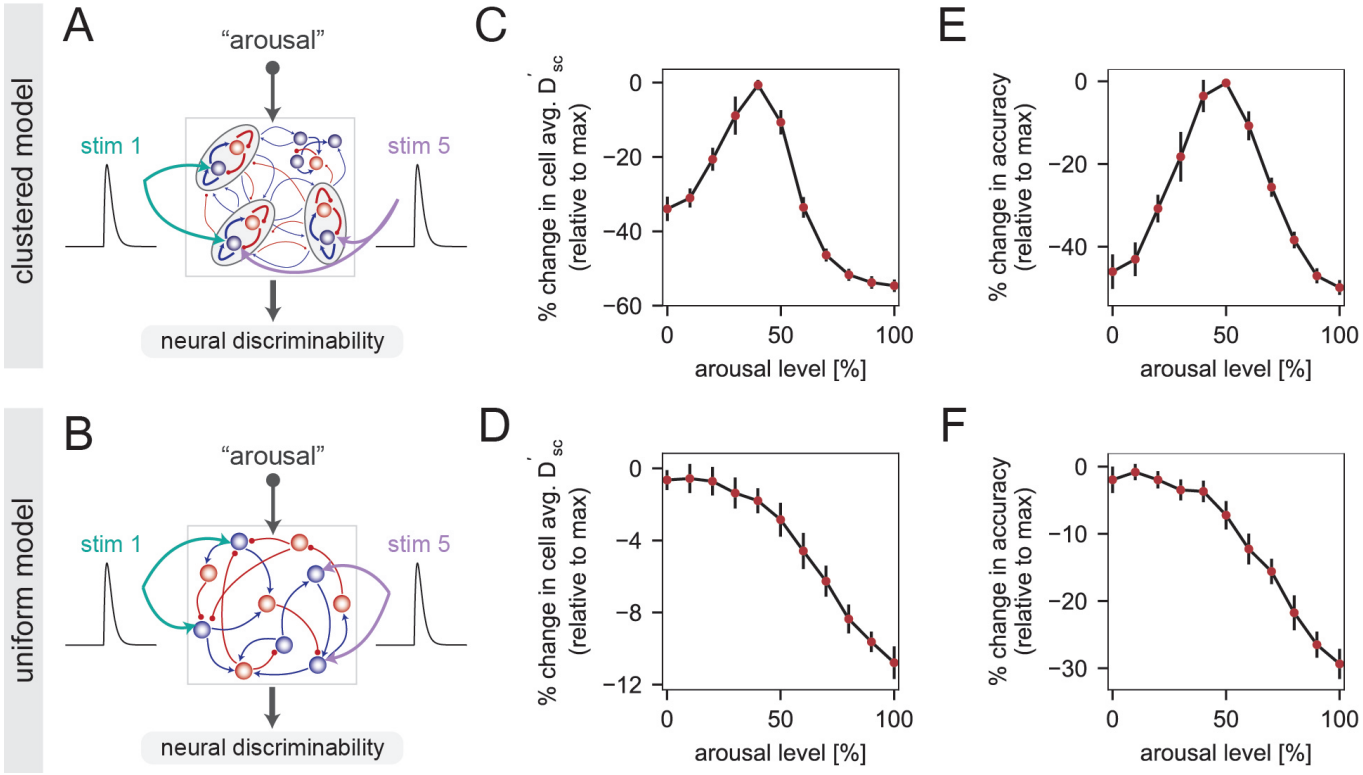

FIG. 5. **The clustered model captures the inverted-U relationship between stimulus discriminability and arousal.** **(A,B)** Five different stimuli were presented several times to the clustered **(A)** and uniform **(B)** networks, and measures of neural stimulus discriminability were computed from the responses. **(C, D)** Percent change in cell-averaged D'_{sc} *vs.* arousal in the clustered (**C**) or uniform (**D**) models (percent change was computed relative to the maximum across arousal levels; Methods). **(E, F)** Same as **(C, D)** but for cross-validated population decoding accuracy (Methods). For this analysis, linear classification was performed using ensemble activity from 10% of the excitatory cells (in the clustered model, an approximately equal number of cells were used from each cluster). In panels **C-F**, data points and error bars indicate the mean ± 1 SD across network realizations. See Fig. S4 for results with different population sizes.

arousal modulates stimulus discriminability, we first used mean-field theory (MFT) to elucidate the effects of arousal on the clustered network's attractor landscape under spontaneous conditions (Methods). Though the MFT does not quantitatively describe the simulations, it provides useful qualitative insights.

At low arousal, MFT reveals the presence of multiple attractors in which different subsets of clusters are highly active (multistable cluster states). Beyond a certain arousal, however, the MFT indicates a transition to a regime with only a single-attractor phase (uniform state), wherein all clusters have the same moderate firing rate (Fig. 6A). The MFT thus predicts that increasing arousal will reduce the contrast between the firing rate of active and inactive firing modes, an intuition that was qualitatively confirmed in network simulations (Fig. 6B). These modulations of the collective dynamics are driven by the suppression of recurrent synaptic excitation onto pyramidal neurons, which hinders the ability of clusters to achieve and maintain a high activity state.

The arousal modulation could also impact the timescale of cluster switching dynamics. To theoretically elucidate the effects, we analyzed a reduced network composed of two excitatory clusters (Fig. 6C Left; Methods). This network also displays cluster states, but has a simplified landscape with two attractors in which either cluster is active (Fig. S5D-H). Using effective mean field theory [24, 25, 69, 70], the attractors can be represented by two potential wells separated by a barrier (Fig. 6C Right). The height h of this barrier controls the rate of stochastic transitions between attractors [21, 24, 71].

The barrier height changed with arousal level. At low arousal, the two attractors were separated by a relatively large barrier, indicating inflexible dynamics with slow switching between attractors (Fig. 6D Left). At intermediate arousal, the two wells were preserved but the barrier height decreased (Fig. 6D Middle), implying more flexible cluster dynamics with faster switching between configurations. For yet larger arousal, there was a transition from the 2-attractor phase to a single-attractor phase (Fig. 6D Right); this transition indicates the loss of metastable cluster states. The theoretical insights from the reduced circuit were verified in simulations of the full clustered model. Specifically, we observed decreases in the average cluster inter-activation interval and activation time with increasing

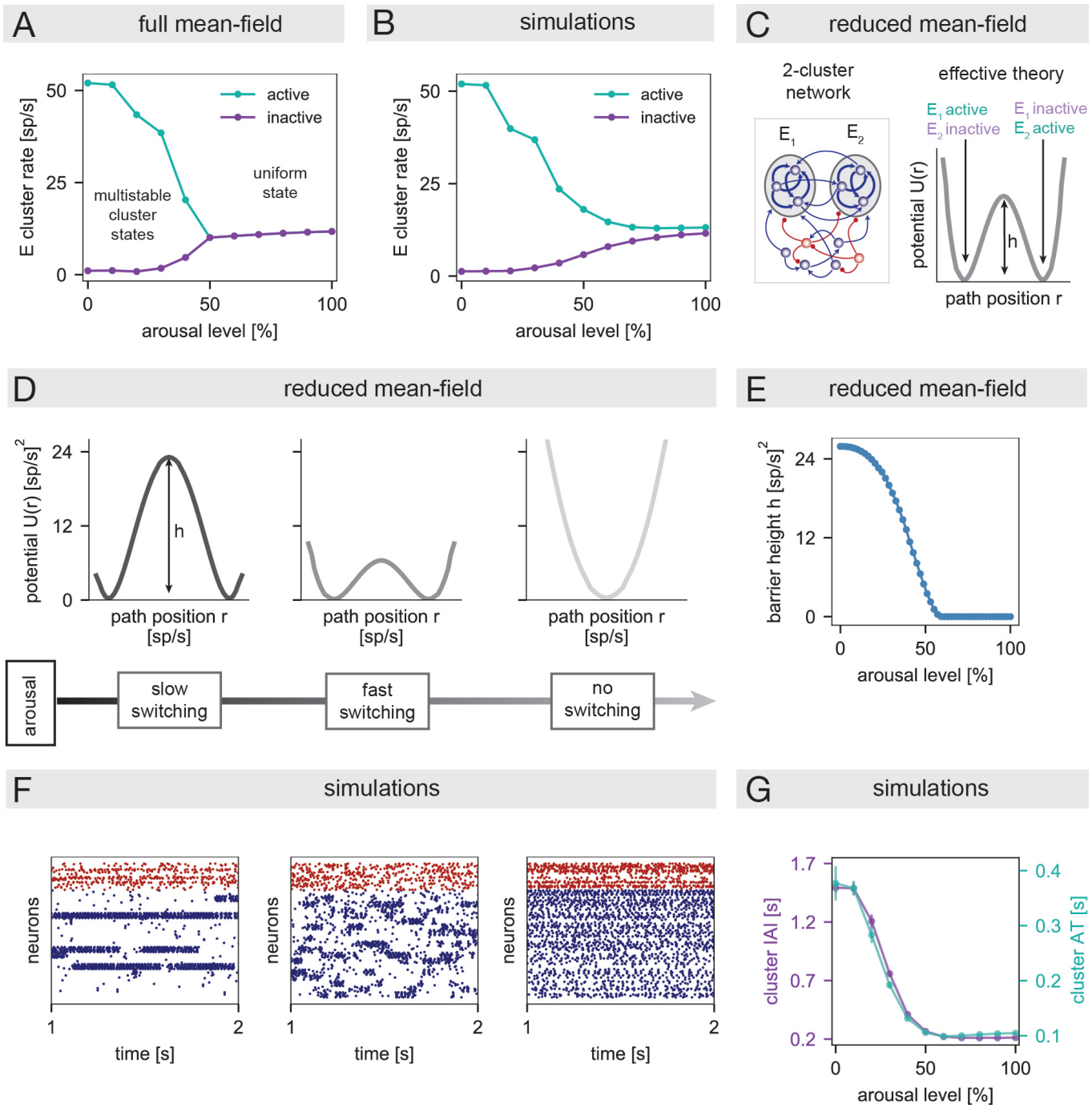

FIG. 6. **The arousal modulation controls the dynamical regime of the clustered model. (A)** Mean-field firing rates of active and inactive excitatory clusters *vs.* arousal. In these analyses, the mean-field calculations used a larger intracluster coupling than the simulations, so the comparison to panel **(B)** is only qualitative (Methods; see also Fig. S5A-C). **(B)** Average firing rate of active and inactive excitatory clusters from simulations *vs.* arousal. **(C)** Schematic of the reduced mean-field analysis using a simplified network of two excitatory clusters. The behavior of the two clusters can be described via an effective potential energy, where the two wells correspond to the network's two attractors (Methods; see also Fig. S5D-H, Table S2). **(D)** The effective potential of the 2-cluster network at three increasing levels of arousal. **(E)** The barrier height h of the effective potential *vs.* arousal. **(F)** Example raster plots from simulations of the full clustered networks at three increasing levels of arousal. **(G)** The average cluster inter-activation interval (IAI, left axis) and average cluster activation time (AT, right axis) *vs.* arousal in simulations of the full clustered networks (Methods). In panels **(B)** and **(G)**, circular markers and error bars indicate the mean ± 1 SD across network realizations.

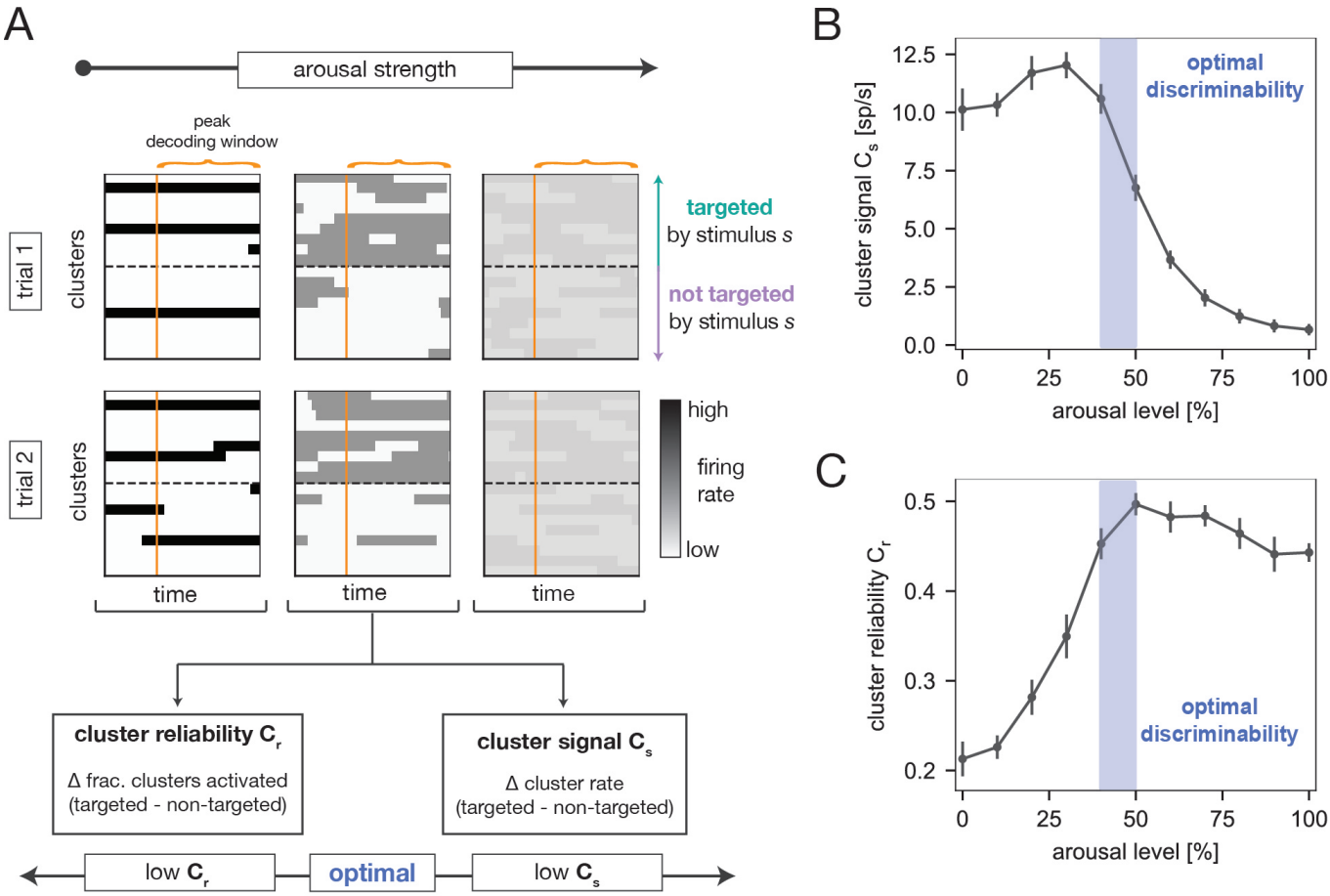

FIG. 7. **Modulations of cluster dynamics underlie the inverted-U relationship.** **(A)** Schematics demonstrating variations in single-trial evoked responses as a function of arousal in the clustered model. Each panel depicts single-trial cluster firing rates in response to a stimulus s (plotted relative to the time of peak decoding accuracy). At a given arousal level, we computed two quantities to characterize the cluster activity pattern: the "cluster signal" C_s and the "cluster reliability" C_r (Methods). **(B)** The cluster signal *vs.* arousal. **(C)** The cluster reliability *vs.* arousal. Data points and error bars indicate the mean ± 1 SD across network realizations. The purple area indicates the region of optimal stimulus discriminability (Fig. 5).

arousal (Fig. 6F Left, Middle; Fig. 6G), in accordance with the shrinking barrier in the reduced network (Fig. 6E). Visual inspection of network activity also revealed a degradation of metastable cluster states with increasing arousal (Fig. 6F Right), consistent with a transition to a uniform phase.

Modulations of cluster dynamics underlie the inverted-U relationship

We used the insights of the previous section to develop intuition for the inverted-U nature of stimulus discriminability. To begin, we note that stimulus identity would be perfectly read-out from population activity if each stimulus could strongly activate all of its targeted clusters on every trial and strongly suppress all non-targeted clusters. To quantify the extent to which the ideal scenario occurs, we examined two properties of network activity following stimulus presentation (Fig. 7A; Methods): (i) the difference between the average firing rates of targeted and non-targeted clusters ("cluster signal"), and (ii) the difference between the fractions of targeted and non-targeted clusters that are in an activated state ("cluster reliability").

The cluster signal increased slightly and then strongly decreased as a function of arousal strength (Fig. 7B). At low arousal, there is a large separation in the spontaneous firing rates of active and inactive clusters; because stimulus presentation biases the activation of targeted clusters, the cluster signal is high in this regime (Fig. 7A Left). When arousal increases to moderate levels, the contrast between active and inactive clusters decreases but remains substantial; at the same time, transitions between cluster states are induced more easily. These effects enable an increase in the relative amount of targeted cluster activation in response to a stimulus, and the cluster signal

remains relatively high (Fig. 7A Middle). As arousal increases further, the spontaneous firing rates of active and inactive clusters further converge. In consequence, the contrast between the average evoked responses of targeted and non-targeted clusters also decreases, and the cluster signal eventually drops to near zero (Fig. 7A Right).

The cluster reliability, in contrast, increased from low to moderate arousal, and then slightly decreased at high arousal (Fig. 7C). At low arousal, spontaneous cluster dynamics are slow and inflexible and only a fraction of all clusters activate in a fixed time window. Because stimuli are not strong enough to completely override the ongoing dynamics, only a fraction of all targeted clusters become activated in response to stimulation, and sometimes non-targeted clusters fail to deactivate. This results in inconsistent activation of targeted clusters and low cluster reliability at low arousal (Fig. 7A Left). As arousal increases, cluster dynamics are more malleable and a larger fraction of clusters activate in a fixed time window. These effects drive the increase in cluster reliability at moderate arousal (Fig. 7A Middle). At very high arousal, non-targeted clusters are not as strongly suppressed (relative to the amount that targeted clusters are activated), which leads to the slight decrease in cluster reliability (Fig. 7A Right); however, it is difficult to estimate reliability at high arousal because the distinction between active and inactive clusters is not well-defined.

The variations in cluster signal and reliability provide intuition for the inverted-U behavior of neural stimulus discriminability. At intermediate arousal, both the signal and reliability are relatively high (Fig. 7B,C). In this optimal regime, stimulus discriminability is maximal. For both lower and higher arousal, either the reliability or signal drop substantially, and discriminability is worse. The key insight is that the arousal modulation affects both the overall strength and consistency of cluster activation patterns, which combine to determine how well the responses to different stimuli can be distinguished.

The clustered model qualitatively captures arousal-related modulations of neural variability

In the clustered model, the transition from the metastable attractor phase to the uniform phase also predicts certain modulations of neural variability. To delineate these effects, we examined the Fano factor (FF), which measures trial-to-trial variability of single-neuron spike-trains (Methods). Increasing arousal was associated with a strong decrease in the FF during spontaneous activity ($\mathrm{FF}_{\mathrm{spont}}$; Fig. 8A). At low arousal, clusters slowly transition between highly active and inactive states (Fig. 6F,G); as a result, single-neuron firing rates strongly fluctuate across time and/or trials, leading to high spike-count variability. As arousal increases, cluster switching dynamics accelerate and the distinction between active and inactive rates decreases; these changes lead to a decrease in spontaneous FF as the model transitions towards the uniform phase. The evoked FF ($\mathrm{FF}_{\mathrm{evoked}}$) also decreased with arousal (Fig. 8B), which is a consequence of stimulus-evoked activity being constrained by the ongoing dynamics. That is, while stimulus presentation does bias activation of targeted clusters, stimuli are not so strong as to be able to activate all targeted clusters simultaneously on every trial (Fig. 7A). The evoked activity thus inherits much of the intrinsic variability present in spontaneous activity. The above-mentioned modulations of cluster dynamics are also reflected in arousal-induced reductions of the coefficient of variation of interspike intervals and low-frequency spike-train power (Fig. S6).

Observed trends in the empirical data qualitatively agreed with predictions from the clustered model. At the population level, the spontaneous FF decreased with pupil diameter (Fig. 8D), and was significantly smaller in states of high pupil-indexed arousal compared to low pupil-indexed arousal (Fig. 8E; Fig. S7A). The population-averaged evoked FF plateaued at moderate-to-large pupil diameters (Fig. 8F), but was still significantly smaller in high arousal compared to low arousal conditions (Fig. 8G; Fig. S7B). Similar trends were also observed for the coefficient of variation of interspike intervals and low-frequency spike-train power (Fig. S6). In sum, the model captures qualitative trends in spike-train variability with arousal, but does not quantitatively match the data in terms of the absolute values of the variability measures.

Past studies indicate that stimulus presentation quenches neural variability [72]. To quantify this effect in our dataset, we estimated the difference between the spontaneous and evoked FF ($\Delta\mathrm{FF} = \mathrm{FF}_{\mathrm{spont}} - \mathrm{FF}_{\mathrm{evoked}}$), marginalized across pupil diameter. Consistent with past reports, we observed a reduction in the FF during evoked conditions (Fig. S7D). Clustered networks were previously proposed to explain this phenomenon [20, 21], and in our model, we also observed some stimulus-induced quenching at low arousal (Fig. 8C). However, the effect is relatively subtle, since our model operates in a regime where stimuli cannot entirely override ongoing activity. The clustered model also exhibited an interaction between arousal and stimulus-induced variability quenching, wherein ΔFF decreased with arousal (Fig. 8C). In the data, we similarly found a decreasing trend in the population-averaged ΔFF with pupil diameter (Fig. 8H), and a significant reduction in ΔFF between highly-constricted and highly-dilated pupil conditions (Fig. 8I; Fig. S7C).

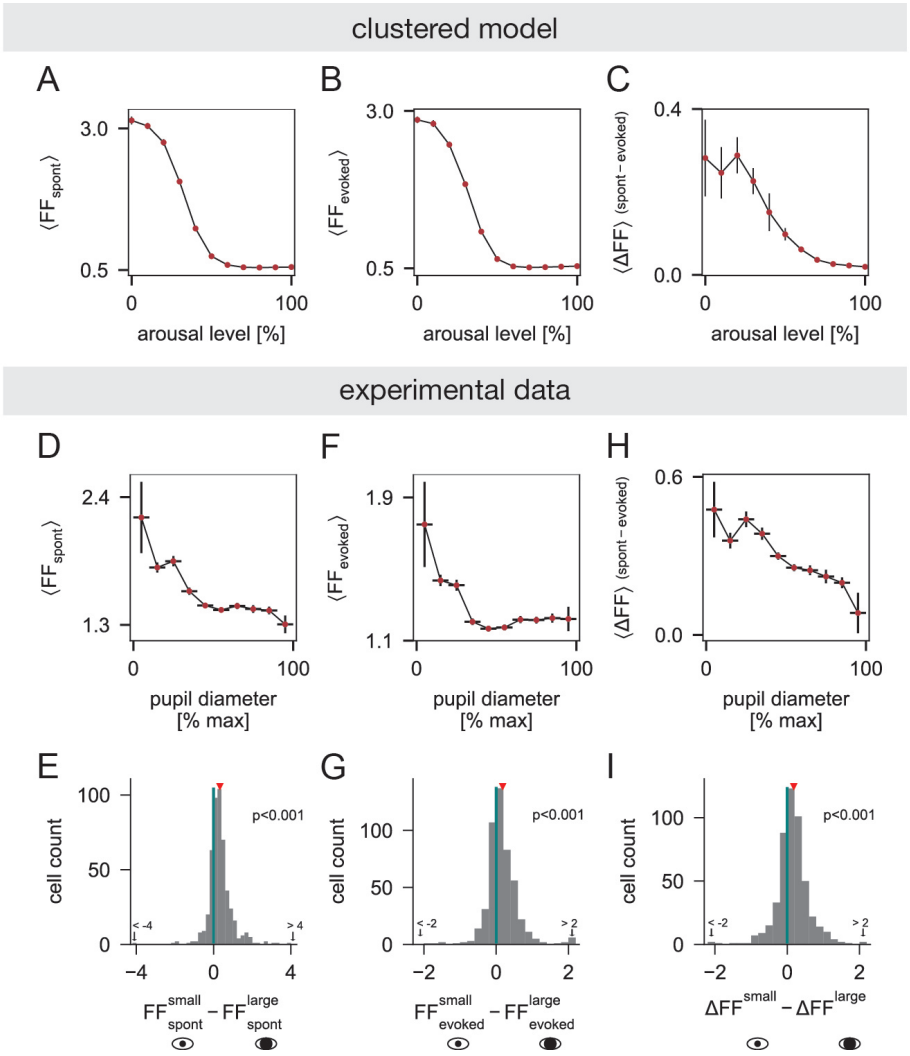

FIG. 8. **The clustered model qualitatively captures arousal-related modulations of neural variability. (A-C)** Results from the clustered model. **(A)** Population-averaged spontaneous FF (FF_{spont}) *vs.* arousal. Data points and error bars indicate the mean ± SD of the population-averaged FF_{spont} across network realizations. **(B)** Same as **(A)** but for evoked FF (FF_{evoked}). **(C)** Same as **(A)** but for spontaneous minus evoked FF (ΔFF). **(D-I)** Results from the experimental data. **(D)** Population-averaged FF_{spont} *vs.* pupil diameter (units pooled over sessions). Horizontal error bars indicate pupil diameter bins; data points and vertical error bars indicate mean ± SEM of FF_{spont} across cells from all sessions that contributed to the corresponding pupil bin. **(E)** Distribution of the difference in FF_{spont} between small and large pupil diameters [$n = 487$ units pooled over 9 sessions with average pupil diameter of smallest (largest) decile bin $\leq 33\%$ ($\geq 67\%$) max dilation; red triangle indicates mean difference; $p < 0.001$, Wilcoxon signed-rank test]. **(F, G)** Same as **(D,E)**, but for FF_{evoked} ($p < 0.001$, Wilcoxon signed-rank test). **(H,I)** Same as **(D,E)**, but for ΔFF ($p < 0.001$, Wilcoxon signed-rank test). See also Figs. S6, S7. Methodological details are provided in Methods.

DISCUSSION

We recorded activity from mouse ACtx during passive tone presentation and found that neural stimulus discriminability followed an inverted-U relationship with pupil-linked arousal. We then showed that this inverted-U relationship could be explained via modulations of metastable attractor dynamics in a clustered network model, with optimal stimulus discriminability achieved near a transition in the dynamical regime of the network. The clustered model further predicted a reduction of neural variability with arousal, which was confirmed in the empirical data. This study thus unifies two different phenomena – arousal-dependent changes in neural discriminability and variability – under the same computational mechanism: modulations of metastable attractor dynamics.

Neural correlates of an inverted-U relationship between performance and arousal in ACtx

Several studies have reported inverted-U relationships between pupil-linked arousal and behavioral performance on auditory tasks [12–16], raising questions about the neural origins of optimal performance at intermediate arousal. Previous work found neural correlates of optimal sound detection in mouse ACtx and medial geniculate nucleus at moderate arousal, as evidenced by reduced variability of spontaneous membrane potential dynamics and increased magnitude and reliability of evoked responses [12]. Here, we found that intermediate arousal was also associated with optimal neural discriminability of tones in mouse ACtx, even in the absence of a task.

Our results suggest that arousal-related modulations of ACtx activity might contribute to the inverted-U relationship between arousal and behavioral performance observed during auditory discrimination tasks [13, 16]. That said, the inverted-U relationship for task performance could also receive contributions from arousal-induced modulations of areas up- or down-stream of ACtx. Important future directions include analyzing task-engaged settings and different stages of the auditory and decision-making pathways [73–75]. It is also important to note that the decoding approach used here assumes that a downstream readout unit acts as an ideal observer implementing linear classification [76]. While a useful benchmark, neural decoding performance can depend on population size, and information readout and perceptual discrimination performance may differ from predictions based on ideal-observer linear decoding frameworks [47]. Finally, we note that one previous study reported a monotonic increase of tone decoding accuracy in mouse ACtx [77]. Several factors could contribute to across-study discrepancies, including differences in recording technique, experimental design, or analysis methods (e.g., type of classifier used). Further examining the conditions under which non-monotonic relationships emerge is an important avenue for future work.

Motivation for the clustered model

We proposed a mechanism for the inverted-U relationship between arousal and neural discriminability using a network model in which neurons were organized into strongly-coupled clusters representing functional neural assemblies (see [19–22, 24, 25].) Our empirical dataset exhibited some evidence of functionally-organized cell ensembles, where we found putative correlation-based clusters with larger tuning similarity than expected by chance alone. Such a relationship between noise correlations and stimulus response similarity is reminiscent of prior work showing positive associations between noise and signal correlations in ACtx [78–81]. More broadly, the clustered model is motivated by evidence of strongly-connected groups of cells in sensory cortices [82–87]. As in the clustered model, neural data indicates that strongly-coupled neurons display similar stimulus responses [83, 85, 87] and that spontaneously-activated cell ensembles are also triggered by stimulation [88, 89], suggesting that cell assemblies may act as basic cortical processing units. The functional architecture of population activity in ACtx, specifically, also appears consistent with the presence of partially-overlapping and strongly-connected neuronal subnetworks [78].

Metastable attractor dynamics are a key feature of the clustered network [20–22, 24, 25], and some analyses have suggested activity patterns reminiscent of those in the model. In particular, Bathellier et al. [48] found that evoked activity patterns in ACtx populations were organized into a small set of discrete "response modes", and that transitions between modes were abrupt, indicative of attractor-like dynamics. Their study also suggested that sounds are represented by the combined activation pattern of several response modes spread across ACtx [49], somewhat akin to the encoding of stimuli via global cluster activation patterns in our circuit model. Other studies in ACtx have observed transient "packets" of elevated spiking activity that occur sporadically during spontaneous periods and that constrain stimulus responses [80, 90], as well as evidence of locally-clustered activity in superficial layers [91]. Although these findings are suggestive of the activity patterns in our network model, more spatially-distributed recordings and targeted perturbation studies are necessary to directly test for the presence of metastable cluster dynamics in ACtx.

Emergence of arousal-induced modulations of neural discriminability and variability in the clustered model

Clustered network models with metastable activity have been used to explain several features of cortical activity and computation [26, 27, 92], including stimulus-induced quenching of variability [20, 21, 23], the emergence of state-sequences during taste processing [22, 93], and context-dependent sensory processing and decision-making dynamics [24, 25, 94]. Previous work examined the response of clustered networks to relatively small parameter perturbations, leading to monotonic variations in stimulus processing efficacy [24, 25]. Here, we explored a broader range of parameter variations, which led to the inverted-U modulation of stimulus discriminability required to explain the data.

In our model, arousal was implemented as a global decrease in synaptic efficacy between excitatory cells and an increase in external excitatory drive. When the clustered network was subjected to this arousal modulation, the inverted-U relationship emerged via a transition from a dynamical phase with slow switching between multiple

metastable attractors to a phase with uniform network activity. Stimulus discriminability was maximized between these two extremes, where stimulus responses were both relatively strong and reliable. Because the transition from the metastable attractor phase to the uniform phase is accompanied by a suppression of slow rate fluctuations, the clustered model also qualitatively captured observed reductions of spiking variability and stimulus-induced variability quenching at high arousal (though see [95] for a study in ferrets where variability quenching was independent of pupil size).

Alternative mechanisms and approaches

We modeled one implementation of arousal that was motivated by empirical evidence and that reproduced the inverted-U relationship. That said, arousal-related neuromodulators affect many biophysical processes to induce diverse alterations of cortical circuit dynamics [54, 55, 58, 59, 62, 64, 65]. For example, neuromodulators can directly excite cortical neurons [96, 97], alter excitability [98], synaptic transmission [61, 66], and cellular firing modes [99], and regulate circuit function via modulation of distinct interneuron classes [100]. The modeling framework presented here could be extended to explore additional mechanisms. Importantly, similar modulations of collective network dynamics can be produced by different circuit perturbations [101]. Even here, we identified an alternative arousal implementation that induced similar functional outcomes to the one studied in the main text (though lacked the same degree of biological plausibility; Fig. S3). Further experimental work is needed to isolate the precise neurophysiological mechanisms of arousal responsible for the inverted-U relationship.

Any circuit mechanism for the inverted-U relationship must incorporate a means of non-linearly modulating the efficacy of stimulus responses. The mechanism we presented here relied critically on the modulation of metastable dynamics in networks with clustering. Importantly, though, our study does not rule out alternative mechanisms that do not require those features (see, e.g., [16, 74]). Along these lines, one particularly relevant study proposed a rate-based decision-making circuit with two excitatory populations, each selective for a different stimulus, and two interneuron classes (VIP and SST), which were both modulated by arousal [16]. In that model, an increase in arousal first improved discrimination performance via VIP-SST-mediated disinhibition of the excitatory pools; but a further increase in arousal saturated the VIP population, leading to a degradation in performance due to SST-mediated inhibition. This study indicates that another way of achieving non-linear performance modulations is to incorporate other types of complexity, such as multiple inhibitory cell types. Other work has shown that cortical sensory processing is impacted by the presence of spontaneously-generated "Up" and "Down" states [50, 102], which are themselves dependent on behavioral state [5] (also discussed further below). Arousal-induced modulation of these global activity patterns could thus also contribute to the inverted-U relationship. Moreover, the right-hand side of the inverted-U relationship might be mediated, at least in part, by motor-related signals, which can suppress tone-evoked activity in ACtx via postsynaptic inhibition of excitatory cells [103]. More generally, a large body of literature demonstrates that information processing capabilities in neural systems are often optimal near criticality [104–109], and some studies have demonstrated enhanced stimulus discriminability at the transition between asynchronous and synchronous dynamics in cortical network models [110, 111]. These studies suggest that other types of dynamical regime transitions – besides the one studied here – may also be able to explain the inverted-U relationship. Future work could seek to further test alternative models and mechanisms and determine which are most consistent with experimental observations.

Mechanisms other than cluster-based switching dynamics could also explain arousal-related decreases of neural variability. In particular, modulations of spontaneous "Up-Down" dynamics – alternating periods of global silence and global activity observed during anesthesia, sleep, and quiet wakefulness [5, 80, 112–115] – could produce similar state-dependent adjustments of single-neuron spiking variability [116–118]. While an advantage of the clustered model is its ability to explain arousal-induced modulations of neural discriminability and variability with a single mechanism, it is possible that both cell assembly dynamics and more distributed Up-Down dynamics contribute to our findings. Indeed, signatures of both have been observed in rodent auditory cortex [48, 49, 78, 80, 90, 102, 114, 116, 119], with prior work suggesting that locally-clustered activity may be more prevalent in superficial layers [49, 91]. In our data, we found evidence of correlated subgroups of neurons suggestive of functionally-meaningful neural clusters. However, we also observed more dispersed correlation structure in the data compared to the model, which could be indicative of more globally-coordinated activity fluctuations. Future experiments with more targeted and spatially-expansive electrode placement (i.e., spanning multiple layers and columns) are needed to isolate and refine the network mechanisms contributing to arousal-related modulations of neural variability and stimulus processing. Building network models that combine global Up-Down fluctuations with clustered neural assemblies [120] may also allow for a more complete description of state-dependent dynamics.

Modality-dependent effects of arousal on sensory-evoked activity

During perceptual decision-making, inverted-U relationships between arousal and performance arise in both auditory [12, 13] and visual [13, 121] tasks. However, relationships between arousal and sensory-evoked activity strongly differ between auditory and visual cortex. In auditory cortex, previous investigations [12] (and this study) report inverted-U relationships between pupil-indexed arousal and measures of stimulus response efficacy, as well as suppressed evoked responses during locomotion [103, 122–124]. In contrast, prior work in visual cortex indicates that sensory-evoked responses are enhanced during high arousal [121, 125] and locomotion [126–129]. The origin of this divergence remains unclear. Although both auditory and visual cortex appear to share some common architectural features (e.g. functional neural assemblies), there may be important variations in their recurrent circuitry or intrinsic neuron properties that lead to differences in ongoing activity, sensory responses, and interplays with arousal. Another possibility is that differences between auditory and visual cortex reflect an underlying distinction in arousal signaling pathways in the two areas. Indeed, the impacts of locomotion in visual cortex are thought to be mediated by a disinhibitory pathway involving VIP interneurons [96], but this is not the case in auditory cortex [123, 124]. Further elucidating the neuromodulatory pathways regulating arousal may help resolve outstanding questions regarding modality-dependent phenomena.

RESOURCE AVAILABILITY

Lead Contact

Requests for further information and resources should be directed to the lead contact, Luca Mazzucato (lmazzuca@uoregon.edu).

Materials Availability

This study did not generate novel reagents.

Data and Code Availability

• The electrophysiological and behavioral data analyzed in this study has been uploaded to the DANDI Archive as Neurodata Without Borders (.nwb) files [130]. The data is publicly available at: https://doi.org/10.48324/dandi.000986/0.251031.1939

• Original code used to run and analyze model simulations, analyze experimental data, and generate manuscript figures has been deposited on Zenodo. The code is publicly available at: https://doi.org/10.5281/zenodo.17497134.

• Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGEMENTS

This work was funded by National Institutes of Health grants R01NS118461 (D.A. McCormick, L. Mazzucato, S. Jaramillo), R35NS097287 (D.A. McCormick.), R01MH127375 and R01DA055439 (L. Mazzucato), R01AG077681 and R01NS127305 (M. Wehr); and National Science Foundation CAREER Award 2238247 (L. Mazzucato)

AUTHOR CONTRIBUTIONS

Conceptualization: L.P., L.M., D.A.M.; methodology: L.P., S.Jo, L.M.; experimental data acquisition: S.Jo., K.Z.; data curation: L.P., S.Ja; investigation: L.P., S.Jo, L.M.; formal analysis, visualization, and software: L.P.; writing

– original draft: L.P., S.Jo, K.Z., L.M.; writing – editing: L.P., K.Z., M.W., S.Ja, D.A.M., L.M.; supervision: S.Ja, D.A.M., L.M.; funding acquisition: M.W., S.Ja, D.A.M., L.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

METHODS

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All procedures were carried out with approval from the University of Oregon Institutional Animal Care and Use Committee. Wild-type animals (2 female mice and 3 male mice ranging between approximately 11-20 weeks at the time of surgery) were of C57BL/6J background purchased from Jackson Laboratory and were bred in-house. Mice were kept on a reverse light cycle and had ad libitum access to food and water. For all analyses, experimental sessions were pooled across both sexes. We did not examine potential influences of sex on our results, which is a limitation of this study.

METHOD DETAILS

Experimental data

Surgical procedures

All surgical procedures were performed in an aseptic environment with mice under 1-2% isoflurane anesthesia, maintaining an oxygen flow rate of 1.5 L/min, and homeothermic maintenance at 36.5 degrees Celsius. Mice were administered systemic analgesia (Meloxicam SR: 4 mg/kg & Buprenorphine SR: 0.5 mg/kg, Wildlife Pharmaceuticals) and a fluid supplement (1 ml lactated ringer's solution) subcutaneously. Fur was removed from the skull, and the skin was disinfected. To access auditory areas, the skin, connective tissue, and part of the right temporalis muscle were resected, and cleaned as necessary. A custom-designed headplate was affixed to the skull using dental cement (RelyX Unicem Aplicap, 3M) and covered with silicone elastomer (Kwik-sil, World Precision Instruments), and skin was affixed to the outside edge of the headpost as necessary (Vetbond, 3M). Mice were allowed to recover for three days in an incubator recovery chamber.

Mice were habituated to handling and head fixation for 2-3 days with increasing duration prior to craniotomy. This was a necessary step for well-being and also helped increase the likelihood that mice entered a broad range of arousal states across the wakefulness spectrum. Habituation to head-fixation atop a treadmill allowed mice to choose to locomote or remain still and quiescent. Craniotomy followed the same aseptic and analgesic procedures as mentioned above. Mice were anesthetized with isoflurane and affixed to the stereotax where a <1 mm circular craniotomy was drilled over the right auditory cortex (AP: -2.9 mm, ML: 4.4 mm relative to bregma) with dura left intact; the above-mentioned coordinates were chosen in an attempt to mainly target primary auditory cortex (A1). A small well was created surrounding the craniotomy with flowable composite (Flow-it, Pentron), and a piece of plastic was secured lateral to the well to act as a shield for the probe. The craniotomy was filled with silicone elastomer (Kwik-sil, World Precision Instruments) until the start of the recording session. Mice were allowed to recover overnight, and recovery was monitored.

Neuropixels recordings

On the day of a recording, a mouse was affixed to a treadmill and the Kwik-sil was removed. The craniotomy was immediately filled with saline, and a high-density silicon probe (Neuropixels 1.0, imec) [131] was inserted perpendicular to the brain surface using a motorized micromanipulator (MP225A, Sutter Instruments) at low speed ($\sim$ 2-4 μm/second) until all layers of the auditory cortex were covered (1.5-2.5 mm). After the Neuropixels probe reached a desired depth, the remaining saline was removed and the craniotomy was filled with 1% agarose mixture in saline and covered with mineral oil to keep the brain surface moist. A recording was started at least 20 minutes after the completion of probe insertion to ensure the stability of the probe and the brain. Recordings were made in up to 5 sessions from one mouse depending on the status of the brain surface. For the last recording session, the Neuropixels probe was covered with DiI (Vybrant solution, Thermofisher Scientific) for histology.

Neurophysiology data was acquired using the PXIe acquisition module (imec) in a NI PXIe-1071 chassis (National Instruments) and OpenEphys software [132] at gain of 250 (LFP), and 500 (APs). An output pulse from the OpenEphys software was manually toggled between 1 Hz and 10 Hz to give an accurate and discrete timestamp to the Power 1401 digitizer, which allowed for accurate alignment and further synchronization of the behavioral data. Neuropixels data was sampled at a rate of 30 kHz. The recorded data was pre-processed with common-average referencing [133],

[134] sorted with Kilosort2 [135, 136], and then manually curated with the phy GUI (https://github.com/cortex-lab/phy). For manual curation, each cluster was compared with other clusters based on the spike waveforms and cross-correlation. Clusters with high similarity were mainly inspected to determine whether they should be merged. Then, the cluster was labeled as a good single unit, multi-units, or noise depending on the quality of the cluster assessed by waveform consistency, amplitude, cross-correlation, and inter-spike intervals. To determine if the good single units were within the auditory cortex, the depth from phy was referenced. In four out of the five mice analyzed, we also verified the recording site of the final session with DiI track spanning after histology. Sessions for which the brain condition was poor, auditory responses were weak, or timestamps could not be aligned were discarded.

Histological Analysis

Following the last recording session, a mouse was anesthetized and perfused using phosphate buffer and 4% paraformaldehyde. Then, the brain was kept in 4% paraformaldehyde, cryo-sectioned (CM3050S, Leica) at 100 μm thickness, and DAPI-stained. Slides were imaged and DiI tracks were manually registered with the Franklin-Paxinos atlas [137].

Spontaneous data and auditory stimulation

Each experimental session consisted of alternating spontaneous and auditory stimulation blocks, repeated for up to 2 hours. During spontaneous blocks, neural activity was recorded in the absence of stimulus presentation; each block lasted for five minutes. A spontaneous block was followed by 25 minutes of auditory stimulation. This design enabled us to record substantial amounts of both spontaneous activity ($\sim$ 20-25 minutes/session) and evoked activity ($\sim$ 75-100 minutes/session). The stimulus set consisted of five pure tones (2, 4, 8, 16, or 32 kHz), which were randomly interleaved and sampled from a uniform distribution. Each tone lasted for 25 ms (cosine ramp-up) followed by a 775 ms inter-stimulus interval (ISI). Auditory stimuli were delivered using custom LabView (National Instruments) scripts. Tones were calibrated to 60 dB SPL and waveforms were generated (NI PXI-4461, National Instruments) at 200 kHz sampling rate, conditioned (ED1, Tucker Davis Technologies), and transduced by electrostatic speakers (ES1, Tucker Davis Technologies).

Behavioral measures

All data collection was conducted using custom LabView scripts. Mice were headfixed atop a cylindrical treadmill (15 cm diameter, 20 cm width) and allowed to freely locomote. Locomotion speed was calculated via a rotary encoder (Encoder Products CO.; 15T-01SF-2500NV1RPP-F03-S1) attached to the axle of the treadmill. Signals from the rotary encoder were continuously converted into cm/s in real-time using LabView software at a rate of 100 Hz, and data was recorded using a Power 1401 digitizer. The running trace was upsampled to match the Neuropixels acquisition rate, and post-hoc analysis was performed using custom python scripts.

The face was lit using an infrared LED (Digi-Key TSHG8200, 830 nm) adjusted to achieve uniform illumination of the face and eye. Additionally, a white LED (RadioShack 5 mm 276-0017) was manually titrated to achieve a wide dynamic range of the pupil, ensuring it remained visible during full dilation. Pupil videos were collected from a camera (Grasshopper 3, FLIR) with a lens (Telecentric TEC-55, Computar) and near-IR Bandpass filter (BN810-43, MidOpt) with FlyCapture software (FLIR). Frames were triggered at 30 Hz through a Power 1401 Digitizer (Cambridge Electronic Design), and online pupillometry was performed using LabView software according to previously described methods [13]. The pupil diameter trace was upsampled to match the Neuropixels acquisition rate, and post-hoc analysis was performed using custom python scripts.

Raw pupil diameter traces were subject to three processing steps: *(1)* artifact removal, *(2)* smoothing, and *(3)* normalization. The pupil-tracking procedure is imperfect, which can lead to artifacts in the pupil diameter traces such as abrupt drops or spikes. To mitigate the effect of these artifacts, we performed both automated and manual cleaning of the pupil traces in each session. Automated artifact removal consisted of finding and discarding periods of time associated with unnaturally-sharp jumps in pupil diameter values between nearby time points. To find these artifacts, we first normalized the pupil diameter trace in a given session by its maximum value. At each time point t_n, we then compared the difference in normalized pupil diameter between t_n and $t_n + 0.5$ ms. If the absolute difference in the normalized pupil diameter between those times exceeded a threshold of 0.08, then we removed the pupil data within a time window starting 250 ms before t_n and ending 500 ms after t_n. This automated procedure removed a large majority of pupil artifacts, but pupil traces were still manually inspected afterwards for outstanding

abnormalities. Remaining problematic time windows were tabulated, and the corresponding pupil data was removed from those periods. Pupil traces were also smoothed after artifact removal for easier manipulation. Smoothing was achieved by taking a moving average of the pupil diameter timecourses using windows of length $1/30^{\text{th}}$ of a second sliding forward in 1 ms steps. After artifact removal and smoothing, the resulting pupil diameter trace for a given session was re-normalized by its maximum value; in all sessions, maximum dilation was associated with movement bouts. This normalization procedure, which has been utilized in several prior studies [11–13, 40, 77, 121], facilitates combining data across sessions. All pupil-based analyses in the main text were performed using the within-session normalized pupil diameter traces. We also verified that the population decoding results (Fig. 2H,I) were robust to an alternative normalization procedure, wherein the pupil traces for each session were all normalized by the *same* value (the largest pupil diameter attained across all sessions combined; Fig. S1F,G). Throughout the text, pupil diameters are reported as a percentage of the maximum value (denoted as "% max dilation").

Periods of time corresponding to artifacts in the pupil diameter trace were also removed from the running trace. The running trace was then smoothed using the same process as for the pupil diameter.

Additional unit selection criteria

After following the procedures described in "*Neuropixels recordings*" to identify putative single units, we implemented some additional criteria for the final unit selection process. First, we discarded all clusters whose average firing rate across the duration of the recording was less than 0.25 spikes/second. The remaining criteria mainly involved further analysis of the spike template amplitudes of each cluster that was identified as "good" after performing the spike sorting and manual curation steps detailed above. Examining the behavior of the template amplitudes (output by Kilosort) for a given cluster across time can reveal potential issues with electrode drift and the general quality of the cluster. Our analysis was designed to search for two potential issues in the spike template amplitudes. First, we considered the shape of the amplitude distribution in a sliding time window, and in each window, we looked for signatures of multiple peaks occurring in the corresponding distribution. The presence of multiple peaks in the amplitude distribution computed from a short block of time is an indication that the particular cluster should not be marked as a well-isolated single unit. Second, we looked for cases when the amplitude appeared to drift towards or away from very low values (i.e., towards or away from the "noise floor") over time. This scenario could suggest that the cluster was not stably-tracked across the recording.

To determine if the distribution of template amplitudes in a short time segment was composed of two or more separate peaks, we examined the amplitude data in non-overlapping, 5-minute windows over the entire dataset. For each window, we used the 'scipy.stats.gaussian_kde' function from SciPy to estimate the probability density function (pdf) of the amplitude data via kernel density estimation with a Gaussian kernel. For each window, we then determined the locations (i.e., amplitude values) and heights of all peaks in the corresponding pdf. If the pdf from a given window had more than one peak, we computed two additional quantities. First, we computed the ratio of the height of the tallest peak to the height of the second tallest peak in the window; we refer to this quantity as the "peak height ratio". Smaller peak height ratios tend to correspond to more even splits of the data between the two groups. Second, we computed the percent difference between the locations of the two highest peaks in a window. Larger percent differences between the peak locations correspond to more well-separated groups. After computing these quantities, we found the set of time windows for which the peak height ratio was less than or equal to ten and for which the percent difference between peak locations was greater than or equal to forty. These cut values were selected so as to find time windows for which there were two (or more) well-separated template amplitude ranges that each contributed substantially to the total amount of data in the window. If at least 10% of all time windows satisfied the above criteria, then the corresponding cluster was not used in subsequent analyses.

To determine if the template amplitude for a given cluster appeared to drift into or out of the "noise floor" over time, we first estimated the noise floor as the smallest template amplitude of the cluster across the whole recording. As above, we then considered the pdf of the amplitudes in 5-minute bins. First, we computed the percent difference between the location of the tallest peak in the pdf of a given window and the location of the noise floor. If this percent difference was less than or equal to fifteen, then the corresponding window was marked as having template amplitudes that were concentrated near the noise floor.

For each window, we also determined the location (i.e., amplitude) of the tallest peak in the pdf. We then computed the smallest and largest of those amplitudes across all time windows, and computed the percent difference between the resulting two values. This quantity, which we refer to as the maximum peak location difference, provides information about the range of template amplitudes sampled across the recording. We removed a cluster from subsequent analyses if the following criteria were met: *(i)* more than 10% (but not all) of time windows either had template amplitudes concentrated near the noise floor or in the bulk, and *(*ii) the maximum peak location difference was greater than or equal to twenty-five. These cut values were chosen so as to try and isolate clusters with significant drift towards

or away from low amplitude values. All analyses in the main text were performed after applying the unit selection procedures described in this section. The number of simultaneously recorded single-units that passed all selection criteria ranged from 30 - 235 per session.

Robustness to cell selection criteria

Because our recordings could have included some cells outside the ACtx, we also tested a more conservative cell selection method that incorporated strict criteria for sound responsiveness. After spike sorting and manual curation (see "*Neuropixels recordings*"), we analyzed the tone-evoked responses of all "good" single units/cells. To begin, evoked data was aligned to stimulus onset ($t = 0$), and a time-dependent firing rate was computed in each trial by counting spikes in a 100 ms window sliding forward in 1 ms steps. The time t of each window was defined as the location of its right edge, and the first time window (baseline activity) was located at $t = 0$ and the last time window was located at $t = 150$ ms. For a given tone, we then compared the firing rate measurements in each evoked time window (i.e., windows with $t > 0$) to the distribution of baseline firing rates ($t = 0$ window) using the Wilcoxon signed-rank test. The p-value for each evoked window was Bonferroni-corrected for multiple comparisons (i.e., using a correction factor equal to the number of evoked windows), and a cell was considered "responsive" to the given tone if at least one of the evoked time windows had a corrected p-value < 0.01. Only cells that responded to at least two tones according to this criteria (and that passed the additional cuts described above in "*Additional unit selection criteria*" were kept for further analysis. We found that the main trends in the neural discriminability, clustering, and neural variability analyses were still evident when using this more conservative cell selection method (Fig. S8).

Network modeling

We modeled a local cortical circuit representing ACtx as a recurrently-connected network of N spiking neurons, N^E of which were excitatory (E) cells and N^I of which were inhibitory (I) cells. Further details on the circuit modeling are provided below. All model parameters are shown in Table S1.

Neural dynamics

Neural activity evolved according to the leaky-integrate-and-fire (LIF) model with exponential excitatory and inhibitory synapses. In this model, the dynamics of the membrane potential of the ith neuron in population $\alpha \in \{E, I\}$ are described by

$$\tau_{\mathrm{m}}^{\alpha} \frac{dV_i^{\alpha}}{dt} = -V_i^{\alpha} + \tau_{\mathrm{m}}^{\alpha} I_{\mathrm{rec},i}^{\alpha} + \tau_{\mathrm{m}}^{\alpha} I_{\mathrm{b},i}^{\alpha} + \tau_{\mathrm{m}}^{\alpha} I_{\mathrm{stim},i}^{\alpha}, \tag{1}$$

where $\tau_{\mathrm{m}}^{\alpha}$ is the membrane time constant of cells in population α, $I_{\mathrm{rec},i}^{\alpha}$ is the recurrent input to cell i in population α from other neurons in the network, $I_{\mathrm{b},i}^{\alpha}$ represents background external input, and $I_{\mathrm{stim},i}^{\alpha}$ is an additional external input representing sensory stimulation. When the membrane potential V_i^{α} reaches a threshold $V_{\mathrm{thresh}}^{\alpha}$, a spike is emitted by the neuron and its membrane potential is reset to a value V_{r}^{α}. After spike emission, the membrane potential remains clamped at the reset value for a refractory period of length $\tau_{\mathrm{ref}}^{\alpha}$.

The recurrent input is a sum of excitatory and inhibitory synaptic currents, such that $I_{\mathrm{rec},i}^{\alpha} = I_{\mathrm{rec},i}^{\alpha E} + I_{\mathrm{rec},i}^{\alpha I}$. These currents obeyed the following differential equations:

$$\tau_{\mathrm{syn}}^{E} \frac{dI_{\mathrm{rec},i}^{\alpha E}}{dt} = -I_{\mathrm{rec},i}^{\alpha E} + \sum_{j=1}^{N_E} W_{ij}^{\alpha E} \sum_k \delta(t - t_j^{k,E}) \tag{2}$$

$$\tau_{\mathrm{syn}}^{I} \frac{dI_{\mathrm{rec},i}^{\alpha I}}{dt} = -I_{\mathrm{rec},i}^{\alpha I} + \sum_{j=1}^{N_I} W_{ij}^{\alpha I} \sum_k \delta(t - t_j^{k,I}). \tag{3}$$

In Eqs. 2 and 3, τ_{syn}^{E} and τ_{syn}^{I} are the excitatory and inhibitory synaptic time constants, and $W_{ij}^{\alpha\beta}$ represents the strength of the synapse from the jth neuron of population $\beta \in \{E, I\}$ to the ith neuron of population α; these weights

depend on the network architecture (see the section "*Recurrent network architectures*" below). Finally, $t_j^{k,\beta}$ is the time of the kth spike emitted by the jth neuron of population β.

In addition to the recurrent input, each neuron in population α received $C_{\text{ext}}^{\alpha E}$ connections from other excitatory cells outside of the local network. The background synaptic input at the ith neuron of population α evolved according to

$$\tau_{\text{syn}}^{E} \frac{dI_{\text{b},i}^{\alpha}}{dt} = -I_{\text{b},i}^{\alpha} + J_{\text{ext}}^{\alpha E} \sum_{j=1}^{C_{\text{ext}}^{\alpha E}} \sum_{k} \delta(t - t_{ij}^{\alpha,k}), \tag{4}$$

where $\mathrm{J}_{\text{ext}}^{\alpha E}$ is the strength of external excitatory synapses to cells in population α, and where $t_{ij}^{\alpha,k}$ is the kth spike time of the jth external cell targeting neuron i in population α. The spike times $t_{ij}^{\alpha,k}$ were generated from a Poisson process with rate $\nu_{\text{ext},i}^{\alpha}$; spike trains were independent for each external synapse to a given cell, and there was no shared input across different cells. The baseline value of the external input rate for cells in population α is denoted as ν_o^{α}.

Finally, sensory stimuli were modeled as smoothly-varying, deterministic external inputs $I_{\text{stim},i}^{\alpha}(t)$ that directly entered the voltage equation of the corresponding neuron. Further details on the stimulus inputs are given in the section "*Sensory stimuli*".

Recurrent network architectures

In the circuit model, the network architecture was either "clustered" or "uniform" (Fig. 3A,B). For the uniform case, neurons of type $\alpha \in \{E, I\}$ received a synaptic connection from $C^{\alpha\beta} = p^{\alpha\beta} N^{\beta}$ randomly chosen neurons of type $\beta \in \{E, I\}$. Moreover, all existing synapses from presynaptic neurons of type β to postsynaptic neurons of type α had the same weight, $J_{\text{U}}^{\alpha\beta}$, in the uniform network. In the clustered model, excitatory and inhibitory neurons were instead arranged into p non-overlapping clusters. Each cluster contained $f^{\alpha} N^{\alpha}$ randomly chosen neurons of type α, and the remaining $(1 - pf^{\alpha})N^{\alpha}$ neurons were placed into an unclustered "background" population. Each neuron in a given cluster of type α received $f^{\beta} C^{\alpha\beta}$ connections from other neurons in the same cluster of type β, $(p-1)f^{\beta} C^{\alpha\beta}$ connections from neurons in different clusters of type β, and $(1-pf^{\beta})C^{\alpha\beta}$ connections from neurons in the background population of type β. Each neuron in the background population of type α received $pf^{\beta} C^{\alpha\beta}$ connections from neurons in clusters of type β and $(1 - pf^{\beta})C^{\alpha\beta}$ connections from other neurons in the background population of type β. In this way, the total number of non-zero synaptic connections was the same for the uniform and clustered networks. The weights of non-zero synaptic connections between neurons within the same cluster, $J_W^{\alpha\beta}$, were generally stronger in magnitude relative to the uniform case ($J_W^{\alpha\beta} = J_+^{\alpha\beta} J_{\text{U}}^{\alpha\beta}$, $J_+^{\alpha\beta} > 1$). Moreover the weights of non-zero synaptic connections between neurons in different clusters, $J_B^{\alpha\beta}$, were generally weaker in magnitude relative to the uniform case ($J_B^{\alpha\beta} = J_-^{\alpha\beta} J_{\text{U}}^{\alpha\beta}$, $0 < J_-^{\alpha\beta} < 1$). Synaptic contacts between cells in the background population and cells in clusters were also weakened relative to the uniform model, and given by $J_B^{\alpha\beta}$. Finally, connection weights between background neurons were unchanged relative to the uniform architecture and equal to $J_{\text{U}}^{\alpha\beta}$.

The uniform and clustered networks were constructed such that the sum of all synaptic weights was the same for the two architectures. For a given value of $J_{\text{U}}^{\alpha\beta}$, this was accomplished by fixing the intracluster weight factor $J_+^{\alpha\beta}$, and solving for the appropriate intercluster weight factor $J_-^{\alpha\beta}$. Following this procedure gives

$$J_-^{\alpha\beta} = \frac{f^{\alpha} + f^{\beta} - pf^{\alpha} f^{\beta} - f^{\alpha} f^{\beta} J_+^{\alpha\beta}}{f^{\alpha} + f^{\beta} - pf^{\alpha} f^{\beta} - f^{\alpha} f^{\beta}}. \tag{5}$$

Sensory stimuli

To model stimulus-evoked activity, sensory signals were incorporated as additional, depolarizing external inputs to the cortical circuit (Eq. 1). For the clustered networks, 50% of the assemblies were chosen at random to receive input from a particular stimulus; for each selected cluster, stimulus-related input was then applied to 50% of its E cells (chosen at random). In this way, two different stimuli in general targeted unique but overlapping sets of clusters. For the uniform networks, a given stimulus was modeled as an external input that was applied to a randomly-selected subset of the E cells; for each stimulus, the total number of stimulated neurons was chosen to be the same as in

the clustered model. Throughout the text, we refer to the cells and/or clusters that receive input from a particular stimulus s as "targeted" by that stimulus, and the cells and/or clusters that do not receive input from stimulus s as "not-targeted" by that stimulus. We presented each model network with five different stimuli, matching the five tones used in the experiments.

If the ith cell of population $\alpha \in \{E, I\}$ was targeted by a given stimulus, then the stimulus-related input to that cell took the form

$$I^{\alpha}_{\text{stim},i}(t) = \begin{cases} 0 & \text{if } t < t_{\text{stim}} \\ A^{\alpha}_{\text{stim}} \times \nu^{\alpha}_{o} C^{\alpha E}_{\text{ext}} J^{\alpha E}_{\text{ext}} \times s(t) & \text{if } t \geq t_{\text{stim}}; \end{cases} \tag{6}$$

otherwise, $I^{\alpha}_{\text{stim},i}(t) = 0 \, \forall t$. In Eq. 6, t_{stim} is the onset time of the stimulus, $A^{\alpha}_{\text{stim}} > 0$ sets the amplitude of the stimulation signal for cells in population α, and $s(t)$ describes the stimulus timecourse. Here, $A^{I}_{\text{stim}} = 0$ since only E cells receive sensory stimulation. For the timecourse $s(t)$, we used a difference of exponentials:

$$s(t) = \gamma[e^{-(t-t_{\text{stim}})/\tau_d} - e^{-(t-t_{\text{stim}})/\tau_r}], \tag{7}$$

where $\gamma = \left[(\tau_r/\tau_d)^{\frac{\tau_r}{\tau_d - \tau_r}} - (\tau_r/\tau_d)^{\frac{\tau_d}{\tau_d - \tau_r}}\right]^{-1}$, τ_r is the rise time constant, and τ_d is the decay time constant.

Arousal modulation

In the model, an increase in arousal was implemented as a simultaneous modulation of two parameters: (i) a decrease in the strength of recurrent synapses between excitatory cells (J^{EE}), and (ii) an increase in the level of external drive to E and I cells ($\nu^{\alpha}_{\text{ext}}$, $\alpha \in \{E, I\}$). The arousal-related reduction in J^{EE} was applied globally to all existing (i.e., all nonzero) E-E synapses: for a pair (i, j) of synaptically-connected E cells, the reduction was modeled by setting $J^{EE}_{ij} = J^{EE}_{ij,o} - \Delta_{J_{EE}} \times J^{EE}_{ij,o}$, where $J^{EE}_{ij,o}$ is the baseline value of the synapse and $\Delta_{J_{EE}} > 0$ is a parameter that sets the strength of the modulation. For a given $x \in [0, 1]$ (where the upper/lower limits of x correspond to arousal levels of 0%/100%), the J^{EE} reduction parameter was given by

$$\Delta_{J_{EE}}(x) = \frac{L}{1 + (x^C - 1)^k}, \tag{8}$$

where $C = \log_2\{1/[1 + (\frac{1-x_o}{x_o})^{1/k}]\}$, $k = 1.25$, $x_o = 0.2$, and $L = 0.75$. As arousal increases from 0 to 100% (i.e., as x increases from 0 to 1), J^{EE} decreases as shown in Fig. 3C(i). The arousal-related increase in $\nu^{\alpha}_{\text{ext}}$ varied from cell-to-cell. For the ith cell in population $\alpha \in \{E, I\}$, the external input modulation was modeled by setting $\nu^{\alpha}_{\text{ext},i} = \nu^{\alpha}_{o} + \Delta^{\alpha}_{\nu_i}$, where ν^{α}_{o} is the baseline value of the external input and $\Delta^{\alpha}_{\nu_i} > 0$ is a cell-dependent parameter that sets the strength of the modulation. For a given $x \in [0, 1]$ (where the upper/lower limits of x correspond to arousal levels of 0%/ 100%), the modulation parameter was given by

$$\Delta^{\alpha}_{\nu_i}(x) = \frac{z^{\alpha}_{i} M}{1 + (x^C - 1)^k}, \tag{9}$$

where $C = \log_2\{1/[1 + (\frac{1-x_o}{x_o})^{1/k}]\}$, $k = 1.25$, $x_o = 0.2$, and $M = 13.125$. For $\alpha \in \{E, I\}$, $z^{\alpha}_{i} \sim \text{Beta}(a, b)$ is a random variable drawn from a Beta distribution with shape parameters $a = 10$ and $b = 10$ (same for both E and I populations). Note that because $z^{\alpha}_{i} \in [0, 1]$, the external drive always increases with arousal, but by varying amounts for different cells (see Fig. 3C(ii) for examples). For the mean-field analysis of the reduced 2-cluster network ("*Effective MFT for reduced 2-cluster networks*"), we used the same arousal model, but multiplied L and M in Eqs. 8 and 9, respectively, by a factor of 0.35 (i.e., we used $L = 0.75 \times 0.35$ and $M = 13.125 \times 0.35$). This adjustment recalibrated the arousal parameters such that they varied over an appropriate range for the 2-cluster model.

Numerical simulations

The dynamical system defined by Eqs. 1-4 was integrated using a discrete time step dt $= 0.5 \times 10^{-4}$ seconds. All spike times were forced to the simulation grid, and exact updates were performed between time steps. At a given level of arousal, we performed simulations on several realizations of the clustered and uniform networks; different network instances were also associated with different realizations of the arousal model (i.e., different random draws of the external input modulations; see Eq. 9 above). For most analyses, we generated 10 realizations of the network architecture, and simulated 30 trials of network activity per stimulus for each network realization. For these simulations, each trial lasted 2.5 seconds and stimulus onset occurred at $t_{\text{stim}} = 1$ second; the pre-stimulus period of each trial was considered "spontaneous" activity. For some analyses, we ran a separate set of simulations to obtain longer continuous blocks of spontaneous activity. For the cluster interactivation and activation timescale analyses ("*Cluster timescales*"; Fig. 6G), we simulated two network realizations, and for each one, we ran 30, 5.2-second-long trials of spontaneous-only activity (no stimulus presentation). For the interspike-interval ("*Variability of interspike intervals*"; Fig. S6A) and power spectra ("*Spectral analyses*"; Fig. S6B,C) analyses, we simulated two network realizations, and for each one, we ran 30, 2.7-second-long trials of spontaneous-only activity. In all simulations, different trials used different random initial conditions for neurons' membrane potentials.

Single-cell discriminability

To examine neural discriminability on a single-cell level, we computed a standard metric for quantifying the separability of two univariate stimulus response distributions. Given the responses of an individual cell to repeated presentations of two stimuli s_A and s_B, one measure of single-cell discriminability (d') is:

$$d'(A,B) = \frac{|\mu_A - \mu_B|}{\sqrt{\frac{1}{2}(\sigma_A^2 + \sigma_B^2)}}, \tag{10}$$

where μ_A and μ_B denote the average responses to the two stimuli, and where σ_A and σ_B denote the standard deviations of the two response distributions.

To compute an overall discriminability index in the model or data, we began by computing timecourses of the single-cell discriminability relative to stimulus presentation. First, trials were parsed according to arousal level (see below) and then aligned to stimulus onset. For each trial of a given stimulus, we then computed binned spike counts of every cell in a sliding window (see subsections below for window parameters used in the model and data). In total, we obtained an array of spike counts (i.e., responses) of dimension $N_{\text{cells}} \times N_{\text{stimuli}} \times N_{\text{trials}} \times N_{\text{time bins}}$. In each time bin, the across-trial mean and standard deviation of the spike counts were used to compute d' for each cell and pair of stimuli, according to Eq. 10. To summarize the discriminability of an individual cell i in time bin t, we computed its average d' over all stimulus pairs, denoted as $\overline{d'}_{i,t}$. We then computed the average across all cells in each time bin, denoted as $\langle \overline{d'}_t \rangle$. An overall discriminability index for the population was defined as the maximum of the timecourse $\langle \overline{d'}_t \rangle$; we denote this index as the cell-averaged D'_{sc}. We also determined the time point $t*$ at which $\langle \overline{d'}_t \rangle$ was maximized, from which we computed an overall discriminability index for each cell i as $D'_{\text{sc},i} = \overline{d'}_{i,t_*}$. In the following two subsections, we provide further details on the single-cell discriminability analyses for the network models and experimental data.

Experimental data

To quantify how arousal level impacted single-cell discriminability in the experimental data, we parsed the trials in a given session according to their pupil diameter. To begin, we computed the average pupil diameter across the pre-stimulus period of each trial (100 ms window preceding tone onset). We then split the trials into ten equally-sized partitions according to the deciles of the pre-stimulus pupil diameter distribution (see Fig. 2A for an example); this partitioning procedure allowed us to use the maximum number of trials for the analysis. Within each decile bin, we also randomly subsampled the trials to ensure that each partition contained the same number of trials per tone frequency.

After parsing the data, we computed the single-cell discriminability separately for each pupil-based partition in a session. For this analysis, spikes from each cell were counted in 100 ms windows incremented in 10 ms steps, and we considered a total time span of 450 ms after stimulus onset. We then computed the cell-averaged and single-cell D'_{sc}

values using the procedure described above.

The arousal-conditioned analysis yielded a single value for the cell-averaged D'_{sc} in each pupil diameter decile of a given session. We now explain the procedure for combining results across sessions in order to understand the aggregate effect of arousal on single-cell discriminability (Fig. 2F). First, for each pupil decile in a given session, we computed the percent change in the cell-averaged D'_{sc} relative to the maximum value of that quantity across all decile bins in the session. We then computed the average pupil diameter of the trials in each decile bin, and binned each data point in a session (one per decile) according to that average pupil diameter. For this binning step, we used ten non-overlapping bins, each of width 10% max-normalized pupil diameter. If more than one data point from the same session fell within a single pupil diameter bin, we stored the average percent change in cell-averaged D'_{sc} across all the data in that bin. This process was then repeated for each session, yielding a collection of data points (percent changes in cell-averaged D'_{sc}) in each pupil diameter bin (gray dots in Fig. 2F). Note that because different sessions explored different pupil dilation ranges, not all sessions contributed to every pupil diameter bin; specifically, there was more data at intermediate diameters relative to very small or large ones. To summarize how single-cell discriminability varied with arousal, we computed the average percent change in cell-averaged D'_{sc} across all sessions in each pupil diameter bin (red curve in Fig. 2F); the spread of the data across sessions in each pupil bin was indicated by a boxplot (Fig. 2F).

To quantitatively test whether single-cell discriminability was improved at intermediate arousal relative to either low or high arousal, we compared the distribution of single-cell D'_{sc} values at intermediate pupil diameter to the distributions at small or large diameters using a paired statistical test. For each session, we first determined the pupil decile whose trials had an average pupil diameter closest to 50% of maximum pupil dilation ("central" decile). We also found the set of sessions for which the trials in the first decile bin had an average pupil diameter $\leq 33\%$ of maximum dilation ("low pupil sessions", LS, 10 sessions in total), and the set of sessions for which the trials in the last pupil decile bin had an average pupil diameter $\geq 67\%$ of maximum dilation ("high pupil sessions", HS, all 15 sessions). To compare D'_{sc} between low and middle pupil diameters, we pooled the single-cell D'_{sc} values from the first decile bin and central decile bin of each low pupil session into two groups: $\{D'_{\mathrm{sc,low\ pupil}}\}_{\mathrm{LS}}$ and $\{D'_{\mathrm{sc,mid\ pupil}}\}_{\mathrm{LS}}$. To compare D'_{sc} between high and middle pupil diameters, we pooled the D'_{sc} values from the last decile bin and central decile bin of each high pupil session into two sets: $\{D'_{\mathrm{sc,high\ pupil}}\}_{\mathrm{HS}}$ and $\{D'_{\mathrm{sc,mid\ pupil}}\}_{\mathrm{HS}}$. We then tested for a significant difference between $\{D'_{\mathrm{sc,low\ pupil}}\}_{\mathrm{LS}}$ and $\{D'_{\mathrm{sc,mid\ pupil}}\}_{\mathrm{LS}}$ (or $\{D'_{\mathrm{sc,high\ pupil}}\}_{\mathrm{HS}}$ and $\{D'_{\mathrm{sc,mid\ pupil}}\}_{\mathrm{HS}}$) using the Wilcoxon signed-rank test. In Fig. 2G, we show the distributions of the differences $\{D'_{\mathrm{sc,mid\ pupil}} - D'_{\mathrm{sc,low\ pupil}}\}_{\mathrm{LS}}$ (top panel) and $\{D'_{\mathrm{sc,mid\ pupil}} - D'_{\mathrm{sc,high\ pupil}}\}_{\mathrm{HS}}$ (bottom panel).

Network models

In the network models, spikes from each cell were counted in 100 ms windows incremented in 20 ms steps along the length of a trial. We then used the previously-described procedure to compute the cell-averaged D'_{sc}. For a given network realization and arousal level, results were based off 30 trials per each of 5 stimuli. To summarize how the overall single-cell discriminability varied with arousal strength, we computed the cell-averaged D'_{sc} at each sampled arousal level for a given network realization. At each arousal, we then computed the percent change in the cell-averaged D'_{sc} relative to the maximum value over all arousal levels. Finally, we averaged the results across ten network realizations to obtain the results in Figs. 5C,D.

Population decoding

Population decoding analyses assess the extent to which stimulus identity can be read-out from single-trial responses of a neural ensemble [138]. Here, we used cross-validated linear classification methods to examine how well stimuli could be discriminated from population activity as a function of arousal. For this analysis, trials were first parsed by arousal and aligned to stimulus onset; the spikes of each cell in the ensemble were then counted in a sliding window moving along the length of a trial. For a given arousal level, this procedure yielded a spike-count array of dimension $N_{\mathrm{cells}} \times N_{\mathrm{trials}} \times N_{\mathrm{windows}}$. The following two subsections below provide details on the data selection and spike-count window parameters for decoding in the experimental sessions and network models.

After obtaining the spike-count array, a linear decoding analysis was performed separately for each time window in a trial. Cross-validated linear classification was implemented using version 0.24.2 of the scikit-learn Python package, and proceeded in several steps. Within a given time window, trials were split into training and testing sets. This was achieved using ten repetitions of stratified, 5-fold cross-validation. By using stratified folds, we ensured that the training and testing sets contained the same proportion of trials per stimulus. For each train-test split (50 in total),

the training data was used to fit a multiclass, linear support vector classifier ('sklearn.svm.LinearSVC' with $C = 0.1$, dual = False and all other parameters set to the defaults). After fitting, the trained model was used to predict the stimulus identity of each trial in the test set.

To assess decoding performance, we computed the average cross-validated classification accuracy. Within a given time bin, the accuracy of a single train-test split was defined as the fraction of test trials whose stimulus identity was correctly predicted. The average cross-validated accuracy of the time window was then computed as the average accuracy across all train-test splits. Repeating this process for each time bin yielded a time-course of cross-validated decoding accuracy relative to stimulus onset. The maximum of this time-course (which we refer to as "peak accuracy" or simply "accuracy") was then computed to summarize the overall decoding performance. Throughout the text, we refer to the time window corresponding to peak decoding accuracy as the "peak decoding window". In the following two subsections, we provide further details on the arousal-conditioned decoding analyses for the experimental data and network models.

Experimental data

In the experimental data, all cells were used as features for the population decoding. To quantify how arousal level impacted decoding performance, trials were grouped according to the deciles of their pre-stimulus pupil diameter distribution, as described in the section "*Single-cell discriminability*". Within each decile bin, we randomly subsampled the trials to ensure that each partition contained the same number of trials per tone frequency. Subsequent decoding analyses were then performed independently for each pupil-based partition of the data. When examining the relationship between decoding performance and arousal in the absence of locomotion, we excluded trials with an average pre-stimulus treadmill speed exceeding 2 cm/sec.

After collecting the relevant subset of data, we computed the spike counts of each cell in every trial using 100 ms windows incremented in 10 ms steps, and we considered a total time span of 600 ms after stimulus onset. We then followed the previously-described procedure to compute the peak decoding accuracy in each pupil decile bin of a session. To combine results across sessions (Fig. 2H), we used the method described in "*Single-cell discriminability*".

To test whether moderate arousal was associated with improvements in population-level decoding, we compared the decoding accuracy at moderate pupil diameters to the accuracy at either small or large diameters. For each session, we first determined the decile whose average pupil diameter was closest to 50% of maximum pupil dilation ("central" decile). We then found the set of "low pupil sessions" (those for which the average pupil diameter of trials in the first decile bin was $\leq 33\%$ of maximum dilation; 10 sessions total) and "high pupil sessions" (those for which the average pupil diameter of trials in the last decile bin was $\geq 67\%$ of maximum dilation; all 15 sessions). For all low pupil (high pupil) sessions, we then tested for a significant difference between the accuracy in the central decile and the accuracy in the first decile (last decile) using the Wilcoxon signed-rank test (Fig. 2I).

Network models

For stimulus decoding in the network models, we sampled a subset of the excitatory cells to be used as features in the classification analysis. In the main text (Fig. 5E,F), decoding was performed using ensembles composed of 10% of the excitatory cell population. For the uniform networks, these ensembles were generated by drawing a random sample of cells from the entire excitatory population. For the clustered networks, we randomly sampled an equal number of cells from each subpopulation (i.e, from each cluster and the background population) until the correct sample size was reached; since the total number cells to be sampled was not evenly divisible by the number of subpopulations, the number of cells remaining after sampling equally from each subpopulation were drawn from a randomly-chosen set of the clusters and/or background population. We also explored the impact of using different ensemble sizes in the decoding analyses (Fig. S4). Specifically, we considered sample sizes of 1, 2, 4, 8, 16, and 32 neurons/subpopulation, corresponding to a total of 19, 38, 76, 152, 304, and 608 features ($\sim$ 1.2%, 2.4%, 4.8%, 9.5%, 19.0% and 38.0% of the excitatory cell population).

For a given cell ensemble, the cross-validated decoding accuracy was computed according to the procedure described in the previous section. For this analysis, we used 100 ms spike-count windows incremented in 20 ms steps along the length of a trial. For a given network realization and arousal level, results were based off 30 trials per each of 5 stimuli, and the decoding accuracy was averaged over 25 different runs, where each run used a different random subsample of cells.

The decoding analysis was performed at several values of arousal for each network realization. For a given network realization, we summarized the impact of arousal by computing the percent change in decoding accuracy at each

arousal level relative to the maximum accuracy obtained across all arousal levels. At each level of arousal, we then computed the average percent change in accuracy across ten different network realizations (Fig. 5E,F).

Relationships between firing rate and arousal

Experimental data

To examine how spontaneous firing rates varied with arousal in the data (Fig. S2A-E), we split the spontaneous periods of each session into smaller windows of length 100 ms. For each window, we computed the spike count of every cell and the average pupil diameter over the window duration. Windows from all spontaneous periods were collected into a single dataset, and were then divided into ten groups according to the deciles of their pupil diameter distribution. For each decile bin, we computed *(i)* the average pupil diameter across all windows in the bin, and *(ii)* the average firing rate of each unit across all windows in the bin (see Fig. S2A,B for examples). Finally, we tested for a monotonic relationship between spontaneous firing rate and arousal by computing the Spearman correlation between a unit's average firing rate in each pupil decile bin and the average pupil diameter in each decile bin (Fig. S2C). A correlation with $p < 0.05$ was considered statistically significant, and the sign of the correlation indicated whether the firing rate of the corresponding unit tended to increase (positive modulation) or decrease (negative modulation) with pupil diameter; non-significant correlations indicated the absence of a clear monotonic relationship between spontaneous firing rate and pupil diameter. We note that the above approach is similar to that used in Christensen and Pillow [139]. Fig. S2D shows the fraction of units (averaged across sessions), with significant positive or negative correlations computed with this method. Results for individual sessions are shown in Fig. S2E.

Network models

To quantify how spontaneous activity was impacted by arousal in the network models, we computed single-cell firing-rates in the absence of sensory stimuli. For a fixed value of arousal, rates of all cells were computed during the 800 ms window preceding stimulus onset; in total, we used 150 trials (5 stimuli $\times$ 30 trials/stimulus) per network realization. We then averaged the spontaneous rates of each neuron across trials, and computed the Spearman correlation between the trial-averaged rate of each cell and the arousal modulation strength. A significant ($p < 0.05$) positive/negative correlation indicated a cell whose firing rate tended to monotonically increase/decrease with arousal strength. Fig. S2F shows the fraction of all neurons in the clustered networks that exhibited significant positive or negative correlations with arousal strength. Similar results for the uniform networks are shown in Fig. S2G.

Determining stimulus-responsiveness

To determine if a cell responded significantly to a particular stimulus, we compared its pre- and post-stimulus activity. Specifically, for each trial of a given stimulus, we computed cell spike counts in the 100 ms window preceding stimulus onset and in the 100 ms window right after stimulus onset. For each cell, the pre- and post-stimulus spike counts were then compared using the Wilcoxon signed-rank test, and the stimulus response was considered significant if the two-sided p-value was < 0.05. In the experimental data, the number of trials per tone (before conditioning on arousal state) ranged from 1076-1517, depending on the session.

Correlation-based clustering analysis

We analyzed noise correlation and tuning similarity matrices to look for evidence of functionally-organized neural clusters in patterns of neural activity (Fig. 4). In what follows, we explain how noise correlations and tuning similarity were computed in the model and data. We then describe the clustering procedure used to extract neural clusters from noise correlation matrices, and the statistical methods employed to test whether the detected clusters were meaningfully organized by tuning similarity.

Noise correlations in the network models

We estimated noise correlations for random samples of cells drawn from either clustered or uniform networks. For a given network realization of the clustered (uniform) model, we subsampled 10% of clustered (all) excitatory cells at random. After subsampling, only cells that responded significantly to at least one stimulus were kept for further analysis (see "*Determining stimulus-responsiveness*" for details). To estimate noise correlations, we computed the spike-counts of each cell in the 100 ms post-stimulus window of every trial. We then computed the Pearson correlation between the spike-count vectors of each pair of cells. To avoid capturing correlations driven by arousal- or stimulus-related changes in firing rate, neuron-by-neuron correlation matrices were computed separately for each stimulus and arousal level. A single, overall correlation matrix was then obtained by averaging across all conditions (i.e., across all stimuli and arousals). For a given network realization, results were based off 30 trials/stimulus/arousal level.

We also generated a set of trial-shuffled correlation matrices for each network. For a given stimulus and arousal level, we randomly and independently permuted the trial spike-count vector of each neuron prior to computing pairwise correlations; a single, trial-shuffled correlation matrix was then obtained by averaging across all stimuli and arousal levels, as above. This independent trial-shuffling of the spike-count vectors destroys correlated variability, and leaves behind only correlations that are expected by chance. We repeated this process 100 times, yielding a set of 100 trial-shuffled correlation matrices for each network realization.

Noise correlations in the experimental data

To estimate noise correlations in the experimental recordings, we began by computing single-cell spike counts in the 100 ms post-stimulus window in every trial. We also computed the average pupil diameter across the 100 ms window preceding stimulus onset, and binned trials according to the deciles of the resulting pre-stimulus pupil diameter distribution. To mitigate the impact of correlations due to arousal- or stimulus-related firing rate modulations, trials were separated by tone and pupil decile, and the same number of trials was subsampled for each combination. Then, for a given tone and pupil bin, the noise correlation between a neuron pair was defined as the Pearson correlation between their spike-count vectors from that stimulus and arousal condition. Finally, an overall estimate of pairwise correlations was obtained by averaging noise correlation matrices over 100 different trial subsamples and over all stimulus and arousal combinations. Only cells that responded significantly to at least one tone were included in the analysis (see "*Determining stimulus-responsiveness*" for details).

For each session, we also generated a set of trial-shuffled correlation matrices. For each tone and pupil bin combination, we randomly and independently permuted each neuron's spike-count vector prior to computing pairwise correlations; a single, trial-shuffled correlation matrix was then obtained by averaging across tones and pupil bins, as above. The shuffling process was repeated 100 times (once for each subsampling of the data), yielding a set of 100 trial-shuffled correlation matrices for each session.

Tuning similarity in the network models

To quantify the similarity between the stimulus responses of two neurons, we computed the correlation between their trial-averaged evoked spike-counts for different stimuli. To begin, we computed a time-dependent spike-count for each cell in every trial of a given stimulus using a 100 ms window sliding forward in 5 ms increments. The single-trial stimulus response of a cell was then defined as its spike-count in the 100 ms window following stimulus onset minus the time-averaged spike-count across the 800 ms period preceding the stimulus. The trial-averaged response was then computed as the mean response across all trials of a given stimulus, aggregated over all arousal levels. Finally, the tuning similarity between a pair of cells was defined as the Pearson correlation between their trial-averaged responses to the five different stimuli.

Tuning similarity in the experimental data

The stimulus tuning similarity between each pair of (tone-responsive) neurons was determined from their trial-averaged stimulus responses. For each cell, the single-trial stimulus response was defined as the difference between the spike-counts in the 100 ms window following stimulus onset and the 100 ms window preceding stimulus onset. The trial-averaged response was then computed as the mean response over all trials of a given stimulus, regardless of

pupil diameter. Finally, the tuning similarity between a pair of cells was defined as the Pearson correlation between their trial-averaged responses to the five different tones.

Hierarchical clustering

After obtaining a noise correlation matrix, we performed a clustering analysis to extract putative neural clusters corresponding to functionally-coordinated groups of cells. Given the noise correlation r_{ij} between cells i and j, we defined the distance between them as $d_{ij} = 1 - r_{ij}$. Hierarchical clustering [140] was then performed on the distance matrix using SciPy routines. In hierarchical clustering, each neuron begins in its own cluster. The pair of clusters that are closest – according to a certain "linkage criteria" – are then merged, and this is repeated until all neurons are in a single cluster. The algorithm thus results in a hierarchical partitioning of cells into clusters, where with N neurons, the lowest level contains N clusters and the highest level contains 1 cluster. Here, hierarchical clustering was executed using 'scipy.cluster.hierarchy.linkage' with the 'average' linkage method, followed by 'scipy.cluster.hierarchy.cut_tree' to obtain the cluster partition at each level.

The number of clusters, k, is a free parameter. To determine the best solution, we defined a measure of partition quality as

$$Q = \langle \overline{r_i^{\mathrm{w}}} - \overline{r_i^{\mathrm{o}}} \rangle, \tag{11}$$

where $\overline{r_i^{\mathrm{w}}}$ is the average correlation between cell i and other cells *within* its cluster (with self-correlation set to zero), $\overline{r_i^{\mathrm{o}}}$ is the average correlation between cell i and cells *outside* its cluster, and $\langle \cdot \rangle$ indicates an average over cells. Computing the partition quality at each hierarchical level results in a curve $Q(k)$; the optimal number of clusters k^* was then defined as $\arg\max_k Q(k)$. The final output of the clustering procedure is a vector that contains the cluster label of each neuron for the partition with k^* clusters.

The clustering algorithm always yields a partition of neurons into clusters. It is therefore important to determine whether the detected clusters are significant relative to surrogate data that does not contain true clusters. To this end, we applied the same clustering procedure to observed and trial-shuffled correlation matrices, and determined the optimal partitions in each case. We then computed a quality measure for each cluster in the optimal partition of the observed (un-shuffled) data, and compared the observed statistic to the distribution derived from the optimal clustering of the trial-shuffled data. For a given cluster c, the cluster quality was defined as

$$Q_c = \langle \overline{r_i^{\mathrm{w}}} - \overline{r_i^{\mathrm{o}}} \rangle_{i \in c}, \tag{12}$$

where $\overline{r_i^{\mathrm{w}}}$ is the average correlation between cell i and other cells *within* its cluster (with self-correlation set to zero), $\overline{r_i^{\mathrm{o}}}$ is the average correlation between cell i and cells *outside* its cluster, and $\langle \cdot \rangle_{i \in c}$ indicates an average over cells in cluster c).

The null distribution of cluster qualities $\{Q_c^{\mathrm{null}}\}$ was generated by computing the quality of each cluster in a given trial-shuffled partition, and then aggregating the values across all shuffled partitions (trivial clusters containing only 1 cell were excluded). For each cluster in the observed data, we then computed a p-value as $p = (1 + b_{\mathrm{null}})/(1 + m_{\mathrm{null}})$, where m_{null} is the total number of clusters in the null distribution, and b_{null} is the number of clusters in the null distribution with a quality Q_c^{null} greater than or equal to the observed statistic Q_c^{obs}. Using the Bonferroni correction for multiple comparisons, detected clusters with $p < 0.05/n_{\mathrm{obs}}$ were considered statistically significant, where n_{obs} is the number of (non-trivial) clusters in the observed data (i.e., the number of comparisons). Fig. 4A,E show examples of observed and trial-shuffled cluster quality distributions for the network models and experimental data.

Cluster-based tuning similarity

The functional relevance of detected clusters was assessed with a permutation test, which quantified whether cells in the same cluster had larger tuning similarity than expected by chance. To begin, we defined a test-statistic, "cluster-based tuning similarity", as

$$G = \langle \overline{s_i^{\mathrm{w}}} - \overline{s_i^{\mathrm{o}}} \rangle_{i \in \{c^*\}}, \tag{13}$$

, where $\overline{s_i^{\mathrm{w}}}$ is the average tuning similarity between cell i and other cells *within* its cluster (with self-similarity set to zero), $\overline{s_i^{\mathrm{o}}}$ is the average tuning similarity between cell i and all cells *outside* its cluster, and $\langle \cdot \rangle_{i \in \{c^*\}}$ indicates an average over cells in significant clusters). In general, G is large when the average tuning similarity between cells in

the same cluster is much greater than between cells in different clusters. To determine if the cluster-based tuning similarity was statistically significant, we compared the value of G computed from the optimal cluster partition ("observed value") to a null distribution obtained by randomly permuting cluster labels across neurons. Specifically, we permuted cluster labels $m_{\text{perm}} = 1000$ times, and computed a p-value as $p = (1 + b_{\text{perm}})/(1 + m_{\text{perm}})$, where b_{perm} is the number of permutations that yield a test statistic as large as the observed value obtained from the optimal cluster labels. The cluster-based tuning similarity was considered significant if $p < 0.05$. Fig. 4D,H show examples of the observed and permuted cluster-based tuning similarity in the clustered network model and experimental data. In the clustered model, the cluster-based tuning similarity was always statistically significant (based on results from 10 cell subsamples from each of 10 different networks).

Validity of the clustering procedure

In the clustered model, ground-truth cluster identities are known. To quantify the accuracy of the clustering procedure, we thus applied it to simulations of the clustered model and computed how well the clusters detected by the algorithm agreed with the ground-truth ones. For a given network realization, we computed the optimal clustering partition as described in the section "*Hierarchical clustering*". To quantify the similarity between the true cluster labels and those predicted by the clustering procedure, we then computed the "adjusted rand score" metric as provided by *scikit-learn*. Using that metric, we found that the clustering algorithm yielded partitions that were $> 99\%$ accurate (average score across 10 random cell subsamples from each of 10 different network realizations.

We also verified that the clustering procedure gave reasonable results when applied to simulations of the uniform network model, which does not have true clusters. For a given network realization, we employed the clustering algorithm and statistical test described in "*Hierarchical clustering*" to determine significant clusters. Using those procedures, we found that only 0.2% of clusters detected in cell ensembles from the uniform model were statistically significant (average over clustering results from 10 random cell subsamples from each of 10 different networks). These results are consistent with the fact that the uniform networks do not have strong clustering.

Mean-field analyses

To obtain theoretical insight into the effects of the arousal modulation on network activity, we performed a series of mean-field analyses for the clustered model. Mean-field theory (MFT) is a commonly-applied technique for studying the collective dynamics of large, recurrently-connected networks of integrate-and-fire neurons [141], and has previously been used to study attractor dynamics in networks of LIF neurons with clusters [19, 22, 24, 25]. In what follows, we first explain the mean-field analysis carried out for the full clustered networks with both excitatory (E) and inhibitory (I) assemblies (associated with Fig. 6A of the main text). We then describe the effective MFT performed on the reduced 2-cluster network (associated with Fig. 6C-E of the main text). Because observed changes in stimulus processing result only from changes in network dynamics induced by the arousal modulation (versus from changes in the stimuli themselves), all mean-field analyses were performed for the "spontaneous" condition (i.e., in the absence of sensory stimulation).

MFT for the full clustered networks

Consider a network of LIF neurons composed of p E clusters, p I clusters, 1 "background" (unclustered) E population, and 1 "background" I population, for a total of $2(p + 1)$ populations. We label the populations with a pair of superscripts (α, γ). The first superscript $\alpha \in \{E, I\}$ labels populations as excitatory or inhibitory, and the second superscript $\gamma \in \{1, ..., p+1\}$ specifies the population number, where the first p indices correspond to the cluster labels and the $p + 1$ index corresponds to the background population. All neurons in the same population described by a specific (α, γ) pair have the same intrinsic parameters and the same recurrent connectivity properties (i.e., receive the same number and strength of inputs from their own and other populations). If the external input parameters were also homogeneous across population (α, γ), then the statistics of the overall input to each cell would be identical and all cells in the population would share the same average steady-state firing rate, $\nu^{\alpha,\gamma}$. The situation is more complex in the simulations, because the level of external input varies from cell-to-cell (due to the quenched randomness in the external drive component of the arousal modulation; see "*Arousal modulation*"). However, for the mean-field analyses, we neglect the heterogeneity in the external input rates $\nu^{\alpha}_{\text{ext},i}$ and assume that all neurons in population (α, γ) are subject to the cell-average external rate $\overline{\nu^{\alpha}_{\text{ext}}} = \frac{1}{N^{\alpha}} \sum_i \nu^{\alpha}_{\text{ext},i}$. Thus, in the mean-field, all neurons in the same population (α, γ) are statistically identical and are described by the same average firing rate $\nu^{\alpha,\gamma}$. Though here

we neglect the quenched variability in the external inputs, we note that it may be possible to incorporate it using an extended mean-field framework [142].

The goal of the mean-field analysis is to solve for the steady-state firing rates of each population. To proceed, one makes a set of assumptions about the operating regime of the network, namely, that each neuron's spike train is described by a stationary Poisson process, that the spike trains of different neurons are independent, and that individual spikes from a presynaptic neuron induce only a small change in the voltage of a postsynpatic neuron relative to it's firing threshold [141]. Under these conditions, one can make the diffusion approximation and replace the presynaptic input to population (α, γ) by a Gaussian white noise with mean $\mu^{\alpha,\gamma}$ and standard deviation $\sigma^{\alpha,\gamma}$. Assuming exponentially-decaying synapses with time constant τ_s, the dynamics of a neuron i in population (α, γ) becomes

$$\tau_m^\alpha \frac{dV_i^{\alpha,\gamma}}{dt} = -V_i^{\alpha,\gamma}(t) + \tau_m^\alpha I_i^{\alpha,\gamma}(t) \tag{14}$$

$$\tau_s \frac{dI_i^{\alpha,\gamma}}{dt} = -I_i^{\alpha,\gamma}(t) + \mu^{\alpha,\gamma} + \sigma^{\alpha,\gamma}\eta_i(t) \tag{15}$$

where τ_m^α is the membrane time constant, $V_i^{\alpha,\gamma}$ is the membrane potential, $I_i^{\alpha,\gamma}(t)$ is the total synaptic input from both external and recurrent sources, and $\eta_i(t)$ is a Guassian white noise obeying $\langle \eta_i(t) \rangle = 0$ and $\langle \eta_i(t)\eta_i(t') \rangle = \delta(t-t')$. The mean $\mu^{\alpha,\gamma}$ and variance $(\sigma^{\alpha,\gamma})^2$ of the input depend on the network architecture. For the clustered networks studied here, we have

$$\mu^{\alpha,\gamma} = \begin{cases} \sum_{\beta=E,I} C^{\alpha\beta} f^\beta J_W^{\alpha\beta} \nu^{\beta,\gamma} + \sum_{\beta=E,I} C^{\alpha\beta} f^\beta J_B^{\alpha\beta} \sum_{\substack{\lambda=1 \\ \lambda\neq\gamma}}^{p} \nu^{\beta,\lambda} + \sum_{\beta=E,I} (1-pf^\beta) C^{\alpha\beta} J_B^{\alpha\beta} \nu^{\beta,p+1} \\ + C_{\text{ext}}^{\alpha E} J_{\text{ext}}^{\alpha E} \overline{\nu_{\text{ext}}^\alpha}, \quad \text{if } \gamma = [1, ..., p] \\ \\ \sum_{\beta=E,I} C^{\alpha\beta} f^\beta J_B^{\alpha\beta} \sum_{\lambda=1}^{p} \nu^{\beta,\lambda} + \sum_{\beta=E,I} (1-pf^\beta) C^{\alpha\beta} J_U^{\alpha\beta} \nu^{\beta,p+1} + C_{\text{ext}}^{\alpha E} J_{\text{ext}}^{\alpha E} \overline{\nu_{\text{ext}}^\alpha}, \quad \text{if } \gamma = p+1 \end{cases} \tag{16}$$

and

$$(\sigma^{\alpha,\gamma})^2 = \begin{cases} \sum_{\beta=E,I} C^{\alpha\beta} f^\beta (J_W^{\alpha\beta})^2 \nu^{\beta,\gamma} + \sum_{\beta=E,I} C^{\alpha\beta} f^\beta (J_B^{\alpha\beta})^2 \sum_{\substack{\lambda=1 \\ \lambda\neq\gamma}}^{p} \nu^{\beta,\lambda} + \sum_{\beta=E,I} (1-pf^\beta) C^{\alpha\beta} (J_B^{\alpha\beta})^2 \nu^{\beta,p+1} \\ + C_{\text{ext}}^{\alpha E} (J_{\text{ext}}^{\alpha E})^2 \overline{\nu_{\text{ext}}^\alpha}, \quad \text{if } \gamma = [1, ..., p] \\ \\ \sum_{\beta=E,I} C^{\alpha\beta} f^\beta (J_B^{\alpha\beta})^2 \sum_{\lambda=1}^{p} \nu^{\beta,\lambda} + \sum_{\beta=E,I} (1-pf^\beta) C^{\alpha\beta} (J_U^{\alpha\beta})^2 \nu^{\beta,p+1} + C_{\text{ext}}^{\alpha E} (J_{\text{ext}}^{\alpha E})^2 \overline{\nu_{\text{ext}}^\alpha}, \quad \text{if } \gamma = p+1 \end{cases} \tag{17}$$

where $\nu^{\beta,\lambda}$ is the firing rate of population (β, λ) with $\beta \in \{E, I\}$, $\lambda \in \{1, ..., p+1\}$; all other parameters in Eqs. 16-17 are defined in the section "*Network modeling*". For each population, μ and σ^2 contain recurrent contributions from the same population and from the other populations in the network, as well as an external contribution from the background drive. The system defined by Eqs. 14- 17, along with the threshold and reset conditions for the membrane potential, can be analyzed using the Fokker-Planck framework [141]. When $\tau_s << \tau_m^\alpha$, the steady-state firing rate of neurons in population (α, γ) satisfies the self-consistent relationship

$$\nu^{\alpha,\gamma} = \Phi^{\alpha,\gamma}[\mu^{\alpha,\gamma}(\boldsymbol{\nu}), \sigma^{\alpha,\gamma}(\boldsymbol{\nu})]. \tag{18}$$

In Eq. 18, $\boldsymbol{\nu} = [\nu^{E,1}, ..., \nu^{E,p+1}, \nu^{I,1}, ..., \nu^{I,p+1}]$ is the vector of firing rates of each population and $\Phi^{\alpha,\gamma}$ is the transfer

function for population (α, γ), given by

$$\Phi^{\alpha,\gamma} = \left[\tau_r + \tau_m^\alpha \sqrt{\pi} \int_{q_r^{\alpha,\gamma}}^{q_t^{\alpha,\gamma}} e^{x^2} \mathrm{erfc}(-x) dx\right]^{-1} \tag{19}$$

where

$$q_r^{\alpha,\gamma} = \frac{V_r^\alpha - \tau_m^\alpha \mu^{\alpha,\gamma}}{\sqrt{\tau_m^\alpha}\sigma^{\alpha,\gamma}} + a\sqrt{\tau_s/\tau_m^\alpha} \tag{20}$$

$$q_t^{\alpha,\gamma} = \frac{V_t^\alpha - \tau_m^\alpha \mu^{\alpha,\gamma}}{\sqrt{\tau_m^\alpha}\sigma^{\alpha,\gamma}} + a\sqrt{\tau_s/\tau_m^\alpha} \tag{21}$$

and with $a = -\zeta(1/2)/\sqrt{2}$ [143].

To find allowed states of the network, we numerically solved the set of $2(p+1)$ self-consistent equations defined by Eq. 18 in conjunction with Eqs. 16 and 17. Importantly, multiple solutions – corresponding to different numbers of active and inactive clusters – can exist for the same set of parameters. In such cases, the solution obtained will depend on the initial guess for the firing rate vector. To systematically deal with this fact, we looked for solutions with n_A active clusters and $p - n_A$ inactive clusters by setting the initial rates for the first n_A E and first n_A I populations to ν^E_{high} and ν^I_{high}, respectively, and the initial rates for the remaining E and I populations to ν^E_{low} and ν^I_{low}, respectively. By choosing $\nu^E_{\text{high}} > \nu^E_{\text{low}}$ and $\nu^I_{\text{high}} > \nu^I_{\text{low}}$ we biased the numerical solver to search for solutions with n_A active clusters; the solution space was then mapped by varying $n_A \in \{0, ..., p\}$.

We denote a self-consistent solution with n_A active clusters as $\boldsymbol{\nu}_{n_A}$. The solution in which all clusters have the same firing rate (i.e., $n_A = 0$) is referred to as the "uniform state" and solutions with $n_A \geq 1$ active clusters are referred to as "cluster states". For cluster states, the n_A active clusters of type $\alpha \in \{E, I\}$ have steady-state rate $\nu^\alpha_{n_A,\uparrow}$ and the $p - n_A$ inactive clusters of type α have rate $\nu^\alpha_{n_A,\downarrow}$, where $\nu^\alpha_{n_A,\uparrow} > \nu^\alpha_{n_A,\downarrow}$. Depending on the network parameters, cluster states may not be found.

Selecting J_+^{EE} for the MFT

To study the impact of the arousal modulation in the MFT, we first examined the effect of the E-to-E intracluster weight factor J_+^{EE}, which controls the dynamical regime of the network [22]. We varied $J_+^{EE} \in [12, 19.5]$ using steps of size $\Delta J_+^{EE} = 0.025$. At each J_+^{EE}, we searched for self-consistent solutions $\boldsymbol{\nu}_{n_A}$ with $n_A \in \{0, ..., 5\}$ active clusters. Whether or not a cluster solution was found for a particular $n_A \geq 1$ depended on the value of J_+^{EE} (Fig. S5A).

To compare to the MFT, we ran an additional set of network simulations in which J_+^{EE} was varied in the range $[12, 19.5]$ in steps of size $\Delta J_+^{EE} = 0.75$. For these simulations, no arousal modulations or sensory stimuli were applied, and we ran 20 trials per each of 5 network realizations; all other parameters were as described in Table S1 and the "*Network modeling*" section. For each simulated trial at a given J_+^{EE}, we computed *(i)* the active cluster rate $\nu^E_{n_A,\uparrow}$ conditioned on a given number of active clusters n_A (see "*Cluster firing rates*"), *(ii)* the probability $P(n_A)$ of finding n_A active clusters (see "*Cluster firing rates*"), and *(iii)* the population average firing rate of all E neurons. Analyses were based on 2.3 seconds of simulated activity per trial, and all quantities were averaged across trials and network realizations. Results are shown in Fig. S5B; note that the active cluster rate $\nu^E_{n_A,\uparrow}$ is only plotted for values of n_A satisfying $P(n_A) \geq 0.2$.

We observed that cluster states emerged at lower values of J_+^{EE} in the simulations compared to the mean-field (Fig. S5A,B). This is potentially due to the finite-size of the simulated networks and the inexact incorporation of synaptic dynamics in the mean-field. Although the mean-field does not quantitatively capture the behavior of the simulations, it can still provide insight into the effects of the arousal modulation. In order to qualitatively compare the theory and simulations as a function of arousal, we considered a fixed intracluster weight factor for the simulations ($J^{EE}_{+,\,\text{sim}}$). We then ran the mean-field at a larger intracluster weight factor $J^{EE}_{+,\,\text{mft}}$, which was chosen to achieve the best match with simulations run at $J^{EE}_{+,\,\text{sim}}$ in the absence of the arousal modulation (i.e, using the baseline network parameters). More specifically, we fixed $J^{EE}_{+,\,\text{sim}} = 15.75$ (default value used throughout the main text), and computed the active cluster rate $\nu^E_{n_A^*,\uparrow,\text{sim}}[J_+^{EE} = 15.75]$ conditioned on the most likely number of active clusters $n_A^* = 3$. In the mean-field, we then determined the value of J_+^{EE} for which the active cluster rate $\nu^E_{n_A^*,\uparrow,\text{mft}}$ most closely matched the value $\nu^E_{n_A^*,\uparrow,\text{sim}}[J_+^{EE} = 15.75]$ from the simulations. This procedure yielded a mean-field intracluster weight factor of

$J^{EE}_{+,\,\mathrm{mft}} = 16.725$ (Fig. S5A), which was then used for the mean-field calculations performed as a function of arousal in the main text (Fig. 6A).

MFT as a function of arousal strength

The mean-field analysis provides the steady-state firing rates of active and inactive clusters, conditioned on a particular number n_A of active clusters. Together, these rates summarize the collective activity patterns of the network. To elucidate how the arousal modulation impacts the dynamics of the clustered networks, we fixed the mean-field E-to-E intracluster weight factor $J^{EE}_{+,\,\mathrm{mft}}$ according to the procedure described in "*Selecting J^{EE}_{+} for the MFT*". We then ran the mean-field analysis as a function of the arousal strength, sampling the same data points as the simulations. In the mean-field, varying the arousal strength impacts the E-to-E synaptic weights (J^{EE}_{W}, J^{EE}_{B}, J^{EE}_{U}) and the mean external inputs to E and I cells ($\overline{\nu^{E}_{\mathrm{ext}}}$, $\overline{\nu^{I}_{\mathrm{ext}}}$). All other network parameters were set to the values given in Table. S1.

For a particular choice of n_A, we solved for the active and inactive mean-field rates, $\boldsymbol{\nu}_{n_A,\uparrow}$ and $\boldsymbol{\nu}_{n_A,\downarrow}$, as a function of arousal strength. The mean-field rates were obtained using the procedure described previously in "*MFT for the full clustered networks*". This process was then repeated for different numbers of active clusters n_A. In general, whether or not a cluster state solution was found for a particular $n_A \geq 1$ depended on the arousal level; beyond a certain arousal strength only the uniform state was found.

Fig. 6A of the main text shows the mean-field rates for active and inactive excitatory clusters as a function of arousal. More specifically, for a given arousal level, the plot shows the cluster rates $\nu^{E}_{n^*_A,\uparrow}$ and $\nu^{E}_{n^*_A,\downarrow}$, where n^*_A was the most frequently observed number of active clusters in the simulations at that arousal level (see section "*Cluster firing rates*"; Fig. S5C). If the cluster state solution was not found for a particular arousal level, then the mean-field rate corresponding to the uniform solution is shown. Note that because the mean-field analysis used a different intracluster weight factor than the simulations ($J^{EE}_{+,\mathrm{mft}} \neq J^{EE}_{+,\mathrm{sim}}$; see "*Selecting J^{EE}_{+} for the MFT*"), the comparison between the mean-field and simulations in Fig. 6 is only meant to be qualitative.

Effective MFT for reduced 2-cluster networks

The MFT described thus far yields the steady-state cluster firing rates, but it cannot make predictions about dynamical transitions between the metastable states. To further understand the switching behavior of the clustered networks (Fig. 6D,E), we adapted the effective MFT developed in [69] and later utilized in [24, 25]. For these calculations, we analyzed a reduced version of the full LIF clustered networks composed of two excitatory clusters E_1 and E_2, one background (unclustered) excitatory population E_b, and one background inhibitory population I_b (Fig. S5D). This 2-cluster network was constructed as described in the section "*Recurrent network architectures*", with the exception that we did not depress inter-cluster weights (see Table S2 for reduced network parameters). With the chosen parameters (and the arousal level set to 0%), the standard MFT predicts the presence of a uniform fixed point and two configurations in which one cluster is active and the other inactive (Fig. S5E). The effective MFT enables insight into dynamical transitions between the two cluster states via a dimensionality reduction process that results in a description of the cluster states as wells in an effective potential energy landscape.

Following Mascaro and Amit [69], the analysis proceeds by splitting the network's populations into two groups: *(i)* a set of "in-focus" populations whose dynamical behaviors are of interest, and *(ii)* a set of "ambient" populations. Here, the two clusters E_1 and E_2 are taken as the in-focus populations, and their rates $\boldsymbol{\nu}^F = (\nu^{\mathrm{E},1}, \nu^{\mathrm{E},2})$ are treated as parameters. The two background populations, E_b and I_b, are considered ambient populations, with rate vector $\boldsymbol{\nu}^A = (\nu^{E,b}, \nu^{I,b})$. For some frozen combination of the in-focus rates $\boldsymbol{\nu}^F = \boldsymbol{\nu}^F_{\mathrm{in}}$, the rates $\boldsymbol{\nu}^A$ of the ambient populations are allowed to adapt, and are computed self-consistently by solving the coupled system of equations

$$\nu^{E,b} = \Phi^{E,b}\Big[\mu^{E,b}(\boldsymbol{\nu}^F_{\mathrm{in}}, \boldsymbol{\nu}^A), \sigma^{E,b}(\boldsymbol{\nu}^F_{\mathrm{in}}, \boldsymbol{\nu}^A)\Big] \tag{22}$$

$$\nu^{I,b} = \Phi^{I,b}\Big[\mu^{I,b}(\boldsymbol{\nu}^F_{\mathrm{in}}, \boldsymbol{\nu}^A), \sigma^{I,b}(\boldsymbol{\nu}^F_{\mathrm{in}}, \boldsymbol{\nu}^A)\Big]. \tag{23}$$

The solution to Eqs. 22 and 23 is denoted as $\boldsymbol{\nu}^A(\boldsymbol{\nu}^F_{\mathrm{in}})$. Feedback from the ambient populations then induces new output rates $\boldsymbol{\nu}^F_{\mathrm{out}} = (\nu^{E,1}_{\mathrm{out}}, \nu^{E,2}_{\mathrm{out}})$ for the in-focus populations, which are given by

$$\nu_{\text{out}}^{E,1} = \Phi^{E,1}\Big[\mu^{E,1}\big(\boldsymbol{\nu}_{\text{in}}^{F}, \boldsymbol{\nu}^{A}(\boldsymbol{\nu}_{\text{in}}^{F})\big), \sigma^{E,1}\big(\boldsymbol{\nu}_{\text{in}}^{F}, \boldsymbol{\nu}^{A}(\boldsymbol{\nu}_{\text{in}}^{F})\big)\Big] = \Phi_{\text{eff}}^{E,1}\Big[\boldsymbol{\nu}_{\text{in}}^{F}\Big] \tag{24}$$

$$\nu_{\text{out}}^{E,2} = \Phi^{E,2}\Big[\mu^{E,2}\big(\boldsymbol{\nu}_{\text{in}}^{F}, \boldsymbol{\nu}^{A}(\boldsymbol{\nu}_{\text{in}}^{F})\big), \sigma^{E,2}\big(\boldsymbol{\nu}_{\text{in}}^{F}, \boldsymbol{\nu}^{A}(\boldsymbol{\nu}_{\text{in}}^{F})\big)\Big] = \Phi_{\text{eff}}^{E,2}\Big[\boldsymbol{\nu}_{\text{in}}^{F}\Big] \tag{25}$$

In Eqs. 22-25, the μ's, σ's, and Φ's are computed similarly to Eqs. 16, 17, and 19, but adjusted for the 2-cluster system.

The induced rates of the in-focus populations, $\boldsymbol{\nu}_{\text{out}}^{F}$, are in general different from the initial rates, $\boldsymbol{\nu}_{\text{in}}^{F}$. By varying $\boldsymbol{\nu}_{\text{in}}^{F}$ and computing the difference $\boldsymbol{\nu}_{\text{out}}^{F} - \boldsymbol{\nu}_{\text{in}}^{F}$ at each point, we obtain a flow map in the $(\nu_{\text{in}}^{E,1}, \nu_{\text{in}}^{E,2})$ plane (see Fig. S5F). This flow map captures the response of the clusters to a particular set of quenched input rates $(\nu_{\text{in}}^{E,1}, \nu_{\text{in}}^{E,2})$, and contains the effect of feedback from the ambient populations. In this way, the map reveals the system's fixed points and the flow of the cluster rates $\nu^{E,1}$ and $\nu^{E,2}$ away from the stationary points. Examination of this reduced 2D description indicates that the two cluster states are attractors of the system, and are linked by an unstable fixed point corresponding to the uniform state ($\nu^{E,1} = \nu^{E,2}$; Fig. S5G).

To understand how the arousal modulation impacts the cluster dynamics, we performed the effective MFT for several values of arousal. For these analyses, we used the same implementation of arousal described previously in the section "*MFT for the full clustered networks*", wherein the quenched randomness in the external inputs is neglected. For each arousal level, we obtained a compact representation of the system by numerically integrating the 2D flow-field along a trajectory connecting the two cluster states via the unstable fixed point (see Fig. S5F). This process results in a 1D effective potential with two wells – corresponding to the two cluster states – separated by a barrier whose maxima corresponds to the uniform state (Fig. S5H; Fig. 6C,D). The height h of this barrier is related to the rate of stochastic transitions between the two cluster states [21, 24, 25, 71]. Computing the barrier height as a function of arousal strength thus provides insight into the effects of the arousal modulation on the cluster dynamics, with lower barriers indicating faster switching and shorter-lived cluster activation periods (Fig. 6E).

Measures of cluster activity in the model

Cluster firing rates

To compute cluster firing rates in the clustered model, we first computed the time-dependent firing rate $r_i(t)$ of each neuron i by convolving its spike train with a Gaussian kernel of width $\sigma = 25$ ms, incremented in 1 ms steps. The firing rate $r_c(t)$ of a given cluster c, was then computed as the average rate of its constituent neurons: $r_c(t) = \langle r_i(t) \rangle_{i \,\in\, \text{cluster } c}$.

To quantify how cluster activity varied with arousal (Fig. 6B) or the intracluster weight factor J_{EE}^{+} (Fig. S5B), we computed active and inactive cluster firing rates during the pre-stimulus period of each trial (here taken as the window spanning [-0.8, -0.1]s relative to stimulus onset). To determine the active and inactive rates, we first computed the time-dependent cluster firing rate $r_c(t)$ of every excitatory cluster in each trial (see section "*Cluster firing rates*"). We then computed the average pre-stimulus cluster firing rate across time and trials, which we denote as $\langle r_c^{\text{base}} \rangle$. In a given trial, a cluster was considered "active" at time t if its baseline-subtracted firing rate, $g_c(t) = r_c(t) - \langle r_c^{\text{base}} \rangle$, was greater than zero (i.e., $g_c(t) > 0$). Given this criteria for cluster activation, we determined the number of active clusters n_A as a function of time during the pre-stimulus period. By pooling across all time points with a particular value of n_A, we then calculated the probability of finding n_A clusters active, as well as the average rate of active and inactive clusters as a function of n_A. We denote the trial-averaged active and inactive cluster firing rates for a given n_A as $r_{n_A,\uparrow}$ and $r_{n_A,\downarrow}$, respectively, and the trial-averaged probability of finding n_A active clusters as $P(n_A)$. We determined the most likely number of active clusters, n_A^*, as the value corresponding to the maximum of the probability $P(n_A)$ (after averaging across network realizations).

At a fixed arousal level, only some values of n_A occur with high likelihood; moreover, the most likely number of active clusters (n_A^*) varies with arousal (see Fig. S5C for the probability of finding n_A active clusters at different arousal levels). To summarize the behavior of the clustered networks as a function of arousal, we thus computed the active and inactive cluster rates conditioned on the most likely number of active clusters at each level of arousal (results shown in Fig. 6B). This analysis was based on 150 trials per network realization (5 stimuli $\times$ 30 trials/stimulus), and results were averaged over 10 different networks. See "*Selecting J_{+}^{EE} for the MFT*" and Fig. S5 for details on the analysis of active and inactive cluster rates as a function of the intracluster weight factor J_{EE}^{+}.

Cluster timescales

To calculate the average cluster interactivation and activation timescales, we analyzed long simulations of spontaneous activity (see the section "*Numerical simulations*" for details). For each trial, we used the threshold criteria described in "*Cluster firing rates*" to determine the time points of cluster activation (discarding the first 0.2 seconds and last 0.1 seconds of each simulation). The cluster interactivation interval and activation time were then calculated as the average duration of all cluster interactivation and activation periods, respectively. Results were then averaged across 30 trials per network realization. Fig. 6G shows the cluster interactivation and activation timescales as a function of arousal (average across two different network realizations).

Cluster signal

To calculate the cluster signal (C_s; Fig. 7B), we first computed the time-dependent firing rate $r_c(t)$ of each excitatory cluster in every trial (see section "*Cluster firing rates*"). From the individual cluster rates, we next computed the average time-dependent rate across all targeted clusters, $r_{\mathrm{T}}(t)$, and the average time-dependent rate across all non-targeted clusters, $r_{\mathrm{N}}(t)$. We then took the difference between the average targeted and non-targeted cluster rates, $\Delta r_{\mathrm{T,N}}(t) = r_{\mathrm{T}}(t) - r_{\mathrm{N}}(t)$, and averaged the difference across the peak decoding window (100 ms window corresponding to peak decoding accuracy; see "*Population decoding*"). This procedure resulted in a single number $\Delta r^*_{\mathrm{T,N}}$ for each trial. The cluster signal was then defined as the average of $\Delta r^*_{\mathrm{T,N}}$ across trials. For each network realization, the cluster signal was computed using 150 trials (5 stimuli $\times$ 30 trials/stimulus); results were then averaged across 10 different network realizations.

Cluster reliability

To compute the cluster reliability (C_r; Fig. 7C), we began by computing the time-dependent, baseline-subtracted rate of each cluster, $g_c(t) = r_c(t) - \langle r_c^{\mathrm{base}} \rangle$ (see section "*Cluster firing rates*"). The baseline-subtracted rate was then used to determine if a given cluster was active during the peak decoding window (see "*Population decoding*") in a given trial. In particular, defining g_c^{W} as the average of $g_c(t)$ over the window of interest, a cluster was considered active during that window if $g_c^{\mathrm{W}} > 0$. Given this criteria, we computed the fractions of targeted and non-targeted clusters, $f_{\mathrm{T}_\uparrow}$ and $f_{\mathrm{N}_\uparrow}$, that were active during the peak decoding window in each trial. We then took the difference between those two fractions, $\Delta f_{\mathrm{T}_\uparrow,\mathrm{N}_\uparrow} = f_{\mathrm{T}_\uparrow} - f_{\mathrm{N}_\uparrow}$, and defined the cluster reliability as the average of $\Delta f_{\mathrm{T}_\uparrow,\mathrm{N}_\uparrow}$ across trials. For each network realization, the cluster reliability was computed using 150 trials (5 stimuli $\times$ 30 trials/stimulus); results were then averaged across 10 different network realizations.

Fano factor analyses

We used the Fano factor to characterize single-cell spiking variability in both the clustered network model and the experimental data. For a given cell, the Fano factor (FF) is defined as

$$\mathrm{FF} = \frac{\mathrm{var}[n_{\mathrm{sp}}]}{\langle n_{\mathrm{sp}} \rangle}, \tag{26}$$

where n_{sp} indicates the spike count of the cell within a fixed time window, and where $\mathrm{var}[\cdot]$ and $\langle \cdot \rangle$ indicate the variance and mean across repeated trials (or observation windows), respectively. In both the model and the data, we computed the FF during both spontaneous and evoked conditions.

Network model

In the clustered network model, FFs were computed across 30 trials/stimulus and then averaged across 5 stimuli for each network realization at a fixed arousal level. For this analysis, we focused on cells in the stimulated excitatory clusters (and in the counterpart inhibitory clusters). Cells that had a low spontaneous rate of < 1 spike/second at any arousal level were excluded from the analysis. To compute the FF of a given cell for a particular stimulus, we

binned the spikes in each trial using a 100 ms window incremented in 20 ms steps. The FF was then computed in each time bin according to Eq. 26, yielding a time course FF(t). The spontaneous FF ($\mathrm{FF}_{\mathrm{spont}}$) was defined as the value of $FF(t)$ in the bin immediately preceding stimulus onset. To summarize the evoked FF, we averaged the single-cell timecourses FF(t) across the relevant set of cells for each stimulus and then across all stimuli; we then determined the time point $t_{\mathrm{FF}_{\mathrm{min}}}$ corresponding to the minimum of the cell- and stimulus-averaged trace. For a given cell and stimulus, the evoked FF, ($\mathrm{FF}_{\mathrm{evoked}}$) was then defined as the value of $FF(t)$ at the time $t_{\mathrm{FF}_{\mathrm{min}}}$. For each cell and stimulus, we also computed the difference between the spontaneous and evoked FFs: $\Delta\mathrm{FF} = \mathrm{FF}_{\mathrm{spont}} - \mathrm{FF}_{\mathrm{evoked}}$. To summarize the results, we averaged each quantity across the relevant set of cells for each stimulus and then across all stimuli; we refer to these overall values as $\langle FF_{\mathrm{spont}}\rangle$, $\langle FF_{\mathrm{evoked}}\rangle$, and $\langle \Delta FF\rangle$. Fig. 8A-C show $\langle \mathrm{FF}_{\mathrm{spont}}\rangle$, $\langle \mathrm{FF}_{\mathrm{evoked}}\rangle$, and $\langle \Delta\mathrm{FF}\rangle$, respectively, as a function of arousal (where red markers and error bars indicate the mean +/− 1 SD across 10 network realizations). Figs. S7H-J also show the impact of varying the spike-count window length on $\langle \mathrm{FF}_{\mathrm{spont}}\rangle$.

Experimental data

To compute evoked FFs as a function of arousal, we computed the average pupil diameter across the 100 ms pre-stimulus period of each trial; trials were then split into ten groups according to the deciles of the pre-stimulus pupil diameter distribution (using the procedure described previously in "*Single-cell discriminability*"). To compute spontaneous FFs, the spontaneous blocks of each session were divided into 100 ms windows, and the average pupil diameter was computed across each one; windows were then grouped by pupil diameter, using the same partitions as for the evoked data. This procedure ensured that spontaneous and evoked Fano factors were evaluated across similar pupil dilation ranges. To account for differing numbers of windows and trials in each pupil decile partition, we subsampled the data such that all pupil partitions contained the same number of windows and trials per tone.

For the spontaneous data, the spike count of each cell was computed in each window within a given pupil-based partition. The FF of each cell i was then computed via Eq. 26, and a final estimate of the spontaneous Fano factor, $FF_{\mathrm{spont},i}$, was obtained by averaging across 100 random subsamples of the data. For the evoked FF, trials were first aligned to stimulus onset. In each trial, spikes from each cell were binned using 100 ms windows incremented in 1 ms steps. Using the trials for a given tone and pupil partition, FFs were calculated in each time bin (up to 150 ms after stimulus onset) according to Eq. 26, and results were averaged across 100 random subsamples of the data. This process yielded a time course $FF_{i,s}(t)$ for each cell i and tone s. To summarize evoked FFs, we averaged the timecourse $FF_{i,s}(t)$ across the tone-responsive cells for a particular stimulus (see "*Determining stimulus-responsiveness*") and then across all stimuli. We then determined the time point $t_{\mathrm{FF}_{\mathrm{min}}}$ corresponding to the minimum of the cell- and stimulus-averaged trace. For a given cell i and tone s, $FF_{\mathrm{evoked},i,s}$ was defined as the value of $FF_{i,s}(t)$ at the time $t_{\mathrm{FF}_{\mathrm{min}}}$. Finally, we obtained an overall evoked FF for cell i – $FF_{\mathrm{evoked},i}$ – by averaging $FF_{\mathrm{evoked},i,s}$ across all tones that induced a significant response in that cell. In each pupil bin, we also computed the difference ΔFF_i between the spontaneous and evoked FFs of cell i: $\Delta FF_i = FF_{\mathrm{spont},i} - FF_{\mathrm{evoked},i}$. Only cells that responded to at least one tone and that had an average spontaneous rate ≥ 1 spike/second in every pupil bin were included in the analyses.

To examine the overall effect of arousal on FF_{spont}, FF_{evoked}, or ΔFF, we aggregated each quantity across cells and sessions in a pupil-dependent manner. For each session, we first calculated the average pupil diameter of the trials in each pupil decile partition; we then binned the FF data in each pupil decile according to the decile's average pupil diameter. For this discretization, we used ten non-overlapping pupil bins, each of width 10% max-normalized pupil diameter. If more than one pupil decile from a given session fell into the same pupil diameter bin, we stored the average value of each data point across those deciles. We then repeated this process for each session, yielding a collection of single-cell FF_{spont}, FF_{evoked}, or ΔFF values in each pupil diameter bin. Note that because different sessions explored different pupil dilation ranges, not all sessions contributed to every pupil diameter bin; specifically, there was more data at intermediate diameters relative to very small or large ones. To summarize how the FF quantities varied with arousal, we computed the average of each quantity across all cells in each pupil diameter bin. We denote these cell-averaged quantities as $\langle FF_{\mathrm{spont}}\rangle$, $\langle FF_{\mathrm{evoked}}\rangle$, and $\langle \Delta FF\rangle$. The results of this analysis are shown in Fig. 8D,F,H, where horizontal error bars indicate pupil diameter bins, and where red markers and vertical error bars indicate the cell-average +/− 1 SEM of each FF quantity (FF_{spont}, FF_{evoked}, or ΔFF) over the cells in each pupil bin. Figs. S7E-J also show the effect of varying the spike-count window length on $\mathrm{FF}_{\mathrm{spont}}$.

To test for a difference in each FF quantity between low and high arousal states, we found the set of sessions that expressed a broad range of pupil diameters. Specifically, we considered sessions for which the average pupil diameter of trials in the first pupil decile was $\leq 33\%$ of maximum dilation and for which the average pupil diameter of trials in the last pupil decile was $\geq 67\%$ of maximum dilation (9 sessions in total). We then pooled the single-cell FF quantities in the first and last pupil decile partition across all sessions with a broad pupil range. This yielded a "small pupil" and "large pupil" dataset for each FF quantity: (i) $\{FF_{\mathrm{spont}}^{\mathrm{small\ pupil}}\}$ and $\{FF_{\mathrm{spont}}^{\mathrm{large\ pupil}}\}$, (ii) $\{FF_{\mathrm{evoked}}^{\mathrm{small\ pupil}}\}$ and $\{FF_{\mathrm{evoked}}^{\mathrm{large\ pupil}}\}$, and (iii) $\{\Delta FF^{\mathrm{small\ pupil}}\}$ and $\{\Delta FF^{\mathrm{large\ pupil}}\}$. For each FF quantity, the

small pupil and large pupil groups were compared using the Wilcoxon signed-rank test. The results are summarized in Fig. 8E,G,I, which show the distributions of the difference in each FF quantity between small and large pupil diameters ($FF_{\text{spont}}^{\text{small pupil}} - FF_{\text{spont}}^{\text{large pupil}}$, $FF_{\text{evoked}}^{\text{small pupil}} - FF_{\text{evoked}}^{\text{large pupil}}$, and $\Delta FF^{\text{small pupil}} - \Delta FF^{\text{large pupil}}$). Fig. S7A-C also show the small and large pupil FF quantities separately (both for single cells and population averages).

To test for overall decreases in neural variability during stimulus presentation relative to spontaneous conditions, we marginalized the data in a session across pupil diameters. Specifically, we combined the evoked trials or spontaneous windows from each pupil decile partition (see above) into two aggregate datasets. Using the aggregate datasets, we then followed the methods described above to compute (i) a pupil-aggregated spontaneous Fano factor $FF_{\text{spont},i}$ of each cell i, and (ii) a pupil-aggregated evoked Fano factor $FF_{\text{evoked},i}$ of each cell i. Only cells that responded to at least one tone and that had an average spontaneous rate of ≥ 1 spk/second were included in the analysis. To test for stimulus-induced variability quenching, we pooled the single-cell FF_{spont} and FF_{evoked} values across all sessions to obtain two groups of data: $\{FF_{\text{evoked}}\}$ and $\{FF_{\text{spont}}\}$. We then compared $\{FF_{\text{evoked}}\}$ and $\{FF_{\text{spont}}\}$ using the Wilcoxon signed-rank test (Fig. S7D).

Variability of interspike intervals

Spike train irregularity was quantified by computing the coefficient of variation of interspike interval (cvISI) distributions during spontaneous activity. Given the interspike-interval (ISI) distribution of a single cell, the cvISI is given by

$$\text{cvISI} = \frac{\sigma_{\text{ISI}}}{\mu_{\text{ISI}}}, \tag{27}$$

where σ_{ISI} and μ_{ISI} are, the standard deviation and mean of the ISIs, respectively [144]. The cvISI is equal to 0 for regular, clock-like spiking and is equal to 1 for Poisson spike trains.

Network model

In the clustered model, single-neuron cvISI values were estimated from several simulations (trials) of spontaneous activity at a fixed level of arousal. First, we computed the ISIs of each cell in a given trial. We then collated the ISIs across all trials into a single distribution (separately for each cell). The spontaneous cvISI of a given cell ($\text{cvISI}_{\text{spont}}$) was then computed from its trial-aggregated ISI distribution. At each arousal level, calculations were based on 30, 2.5 second-long trials of spontaneous activity. Results were summarized as the average $\text{cvISI}_{\text{spont}}$ across all clustered cells that had an average firing rate of at least 1 spike/second at each arousal level. We refer to this population averaged quantity as $\langle \text{cvISI}_{\text{spont}} \rangle$. Fig. S6A shows $\langle \text{cvISI}_{\text{spont}} \rangle$ as a function of arousal, where each data point indicates the mean across 2 network realizations.

Experimental data

To compute cvISIs in the neural data, the spontaneous blocks of each session were split into 2.5 second-long windows, and the average pupil diameter was computed across each one. In order to obtain similar arousal-based partitions of the data across different analyses, the spontaneous windows were then grouped – based on their pupil diameter – into one of the ten decile partitions of the pre-stimulus pupil diameter distribution (same splits used for the neural discriminabilty analysis; see "*Single-cell discriminability*"). This discretization allowed us to evaluate changes in the spontaneous cvISI ($\text{cvISI}_{\text{spont}}$) across a broad range of arousal states while maintaining several trials in each pupil decile split. Since different pupil diameter partitions contained different amounts of data, we subsampled the same number of windows from each partition. For a given pupil-based split of the data, we then computed the ISIs of each cell in every time window. The single-unit ISIs were then combined across all time windows in the given pupil partition, and the $\text{cvISI}_{\text{spont}}$ of each cell was computed from its trial-aggregated ISI distribution. This was repeated for 100 different random subsamplings of the data, and a final estimate of $\text{cvISI}_{\text{spont}}$ in each pupil partition was computed as the average across subsamples. Only cells that had an average spontaneous firing rate of at least 1 spike/second in all pupil partitions were included.

To examine the average behavior of $\text{cvISI}_{\text{spont}}$ as a function of arousal, we combined the data over cells and sessions in a pupil-dependent manner, using the method described in "*Fano factor analyses*". The result is shown

in Fig. S6D, where horizontal error bars indicate the pupil diameter bins, and where red markers and vertical error bars indicate the average $\langle \text{cvISI}_{\text{spont}} \rangle$ +/− 1 SEM across all cells in each pupil bin. To quantitatively compare $\text{cvISI}_{\text{spont}}$ between low and high arousal conditions, we used the procedure described previously for the FF quantities (see "*Fano factor analyses*"). First, we found all sessions that expressed a broad range of pupil diameters. We then pooled the single-cell $\text{cvISI}_{\text{spont}}$ values in the first and last pupil decile partition across all of those sessions (9 in total), yielding two groups of data: $\{\text{cvISI}_{\text{spont}}^{\text{small pupil}}\}$ and $\{\text{cvISI}_{\text{spont}}^{\text{large pupil}}\}$. Finally, the small pupil and large pupil groups were compared using the Wilcoxon signed-rank test, and results were visualized by plotting the distribution of the difference $\text{cvISI}_{\text{spont}}^{\text{small pupil}} - \text{cvISI}_{\text{spont}}^{\text{large pupil}}$ (Fig. S6E). Fig. S6F,G also show $\text{cvISI}_{\text{spont}}^{\text{small pupil}}$ and $\text{cvISI}_{\text{spont}}^{\text{large pupil}}$ separately (for both single cells and population averages).

Spectral analyses

We utilized spectral analyses to characterize the temporal structure of spike trains during spontaneous periods in both the network model (Fig. S6B,C) and the experimental data (Fig. S6H-L). To compute the power spectrum of a neuronal spike train from a single trial (time window) of length T, we first binned the spike train at a fine temporal resolution of $\Delta t = 1$ ms. The power spectrum of the binned spike train was then estimated using the multitaper method applied to point processes, as described in [145] and numerically-implemented in [146]. For the multitaper estimates, we used a time-bandwidth product of $TW = 2$ and averaged over $2TW - 1 = 3$ tapers. The multitaper estimate of the spectrum from a given trial was then normalized by the average firing rate of the neuron across that trial; this rate-normalization is equivalent to normalizing the spectrum by that of a Poisson process with the same firing rate. Normalized spectra for a given neuron were then averaged across all trials of a particular condition to obtain a final, normalized power spectrum $S_{\text{norm}}(f)$. The low-frequency power was computed as the average of $S_{\text{norm}}(f)$ between 1-4 Hz.

Network model

In the clustered network model, single-neuron spectra were estimated from several simulated trials of spontaneous activity conditioned on a particular value of arousal. Specifically, for a given network realization and arousal level, we used the method described above to compute the normalized spectrum $S_{\text{norm},i}(f)$ and the low-frequency power $P_{\text{spont},i}^{\text{L}}$ of cell i; these calculations were based on 30, 2.5 second trials of spontaneous activity. To summarize the overall extent of low-frequency fluctuations, we computed the average low-frequency power across all clustered cells that had a firing rate of at least 1 spike/second for all arousal levels. We refer to this cell-averaged low-frequency power as $\langle P_{\text{L,spont}} \rangle$. Fig. S6B shows the population-averaged power spectrum (rate-normalized) and Fig. S6C shows $\langle P_{\text{L,spont}} \rangle$ as function of arousal (averages across 2 network realizations).

Experimental data

To compute power spectra in the experimental data, the spontaneous blocks of each session were divided into 2.5 second-long windows. The windows were then grouped according to pupil diameter using the same scheme that was implemented for the cvISI analysis (see "*Variability of interspike intervals*"). Before computing the power spectra in a given session, we subsampled the same number of windows from each pupil decile partition (we used the maximum possible number of windows given the distribution across pupil partitions); results were then averaged across 50 random subsamplings. For these analyses, we only included cells that had an average spontaneous firing rate ≥ 1 spike/second in all pupil decile partitions, and that had a non-zero spike count in all sampled time windows.

For each pupil decile partition in a session, we computed the normalized spectrum $S_{\text{norm}}(f)$ and low-frequency power $P_{\text{L,spont}}$ of each cell. To examine overall trends in $P_{\text{L,spont}}$ as a function of arousal, we aggregated the data over cells and sessions in a pupil-dependent manner, using the method described in "*Fano factor analyses*". The result is shown in Fig. S6I, where horizontal error bars indicate the pupil diameter bins, and where red markers and vertical error bars indicate the cell-average low-frequency power $\langle P_{\text{L,spont}} \rangle$ +/− 1 SEM in each pupil bin. To compute the cell- and session-averaged power spectrum as a function of arousal (Fig. S6H), we followed the same procedure, but used three large pupil bins for combining data across sessions ($[0-33)\%$, $[33-67)\%$, and $[67-100]\%$ of maximum dilation).

To test for overall changes in low-frequency power between low and high arousal states, we used the method described previously for the FF quantities (see "*Fano factor analyses*"). First, we found all sessions that expressed a broad range of pupil diameters. We then pooled the single-cell $P_{\text{L,spont}}$ values in the first and last pupil decile partition

across all of those sessions (9 in total), yielding two groups of data: $\{P_{\text{L,spont}}^{\text{small pupil}}\}$ and $\{P_{\text{L,spont}}^{\text{large pupil}}\}$. Finally, the small pupil and large pupil groups were compared using the Wilcoxon signed-rank test, and results were visualized by plotting the distribution of the difference $P_{\text{L,spont}}^{\text{small pupil}} - P_{\text{L,spont}}^{\text{large pupil}}$ (Fig. S6J). Fig. S6K,L also show individual values of $P_{\text{L,spont}}^{\text{small pupil}}$ and $P_{\text{L,spont}}^{\text{large pupil}}$ (for both single cells and population averages).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data analysis and simulations were carried out in Python, and utilized the NumPy package [147], SciPy library [148], scikit-learn library [149], and Nitime library [150]. Details of the statistical tests used in this study are provided in the figure legends and/or relevant sections of the Methods. For all statistical comparisons of two paired samples, we used the non-parametric Wilcoxon signed-rank test (implemented with the 'scipy.stats' module from SciPy). For these analyses, "n" refers to either the number of trials, number of cells/units, or the number of sessions (as indicated for each analysis), and significance was determined based on the two-sided p-value. For the clustering analysis, cluster significance was assessed by comparing the observed cluster quality against the distribution obtained under a trial-shuffled null model, and the significance of the cluster-based tuning similarity was assessed with a permutation test (see "*Correlation-based clustering analysis*" for details). When relevant, significance levels were corrected for multiple comparisons using the Bonferroni correction. In cases where some data was excluded from a statistical test (e.g., based on firing rate thresholds), the criteria is specified in the relevant Methods sections. The meaning of error bars (either SEM or SD) is indicated in each figure caption. Boxplots display the median and the first and third quartiles of the data, with the whiskers extending from the quartiles to $\pm$ 1.5 of the interquartile range. Violin plots display probability densities, and tick marks indicate the minimum, maximum, and median of the data.

[1] 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.

[2] David A McCormick, Dennis B Nestvogel, and Biyu J He. Neuromodulation of brain state and behavior. *Annual review of neuroscience*, 43:391–415, 2020.

[3] Seung-Hee Lee and Yang Dan. Neuromodulation of brain states. *neuron*, 76(1):209–222, 2012.

[4] Steven W Flavell, Nadine Gogolla, Matthew Lovett-Barron, and Moriel Zelikowsky. The emergence and influence of internal states. *Neuron*, 2022.

[5] Kenneth D Harris and Alexander Thiele. Cortical state and attention. *Nature reviews neuroscience*, 12(9):509–523, 2011.

[6] Laura Busse, Jessica A Cardin, M Eugenia Chiappe, Michael M Halassa, Matthew J McGinley, Takayuki Yamashita, and Aman B Saleem. Sensation during active behaviors. *Journal of Neuroscience*, 37(45):10826–10834, 2017.

[7] Gary Aston-Jones and Jonathan D Cohen. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28:403–450, 2005.

[8] Robert Mearns Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology & Psychology*, 18:459–482, 1908.

[9] Sebastiaan Mathôt. Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), 2018.

[10] 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):13289, 2016.

[11] Lindsay Collins, John Francis, Brett Emanuel, and David A McCormick. Cholinergic and noradrenergic axonal activity contains a behavioral-state signal that is coordinated across the dorsal cortex. *Elife*, 12:e81826, 2023.

[12] Matthew J McGinley, Stephen V David, and David A McCormick. Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron*, 87(1):179–192, 2015.

[13] Daniel Hulsey, Kevin Zumwalt, Luca Mazzucato, David A McCormick, and Santiago Jaramillo. Decision-making dynamics are predicted by arousal and uninstructed movements. *Cell Reports*, 43(2), 2024.

[14] Peter R Murphy, Ian H Robertson, Joshua H Balsters, and Redmond G O'connell. Pupillometry and p3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, 48(11):1532–1543, 2011.

[15] Leonhard Waschke, Sarah Tune, and Jonas Obleser. Local cortical desynchronization and pupil-linked arousal differentially shape brain states for optimal sensory performance. *Elife*, 8:e51501, 2019.

[16] Lola Beerendonk, Jorge F Mejías, Stijn A Nuiten, Jan Willem de Gee, Johannes J Fahrenfort, and Simon van Gaal. A disinhibitory circuit mechanism explains a general principle of peak performance during mid-level arousal. *Proceedings of the National Academy of Sciences*, 121(5):e2312898121, 2024.

[17] Brian J Schriver, Svetlana Bagdasarov, and Qi Wang. Pupil-linked arousal modulates behavior in rats performing a whisker deflection direction discrimination task. *Journal of neurophysiology*, 120(4):1655–1670, 2018.

[18] R Becket Ebitz, John M Pearson, and Michael L Platt. Pupil size and social vigilance in rhesus macaques. *Frontiers in neuroscience*, 8:100, 2014.

[19] Daniel J Amit and Nicolas Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 7(3):237–252, 1997.

[20] Gustavo Deco and Etienne Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS computational biology*, 8(3):e1002395, 2012.

[21] Ashok Litwin-Kumar and Brent Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature neuroscience*, 15(11):1498–1505, 2012.

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

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

[24] Luca Mazzucato, Giancarlo La Camera, and Alfredo Fontanini. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nature neuroscience*, 22(5):787–796, 2019.

[25] David Wyrick and Luca Mazzucato. State-dependent regulation of cortical processing speed via gain modulation. *Journal of Neuroscience*, 41(18):3988–4005, 2021.

[26] Braden AW Brinkman, Han Yan, Arianna Maffei, Il Memming Park, Alfredo Fontanini, Jin Wang, and Giancarlo La Camera. Metastable dynamics of neural circuits and networks. *Applied Physics Reviews*, 9(1), 2022.

[27] Giancarlo La Camera, Alfredo Fontanini, and Luca Mazzucato. Cortical computations via metastable activity. *Current opinion in neurobiology*, 58:37–45, 2019.

[28] Ziv Gil, Barry W Connors, and Yael Amitai. Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, 19(3):679–686, 1997.

[29] Morgana Favero, Gladis Varghese, and Manuel A Castro-Alamancos. The state of somatosensory cortex during neuromodulation. *Journal of neurophysiology*, 108(4):1010–1024, 2012.

[30] Robert B Levy, Alex D Reyes, and Chiye Aoki. Nicotinic and muscarinic reduction of unitary excitatory postsynaptic potentials in sensory cortex; dual intracellular recording in vitro. *Journal of neurophysiology*, 95(4):2155–2166, 2006.

[31] Candace Y Hsieh, Scott J Cruikshank, and Raju Metherate. Differential modulation of auditory thalamocortical and intracortical synaptic transmission by cholinergic agonist. *Brain research*, 880(1-2):51–64, 2000.
[32] Emmanuel Eggermann and Dirk Feldmeyer. Cholinergic filtering in the recurrent excitatory microcircuit of cortical layer 4. *Proceedings of the National Academy of Sciences*, 106(28):11753–11758, 2009.
[33] Fumitaka Kimura and Robert W Baughman. Distinct muscarinic receptor subtypes suppress excitatory and inhibitory synaptic responses in cortical neurons. *Journal of neurophysiology*, 77(2):709–716, 1997.
[34] Lu Dinh, Tram Nguyen, Humberto Salgado, and Marco Atzori. Norepinephrine homogeneously inhibits α-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate-(ampar-) mediated currents in all layers of the temporal cortex of the rat. *Neurochemical research*, 34:1896–1906, 2009.
[35] Minoru Ohshima, Chiaki Itami, and Fumitaka Kimura. The α2a-adrenoceptor suppresses excitatory synaptic transmission to both excitatory and inhibitory neurons in layer 4 barrel cortex. *The Journal of Physiology*, 595(22):6923–6937, 2017.
[36] Masayuki Kobayashi, Masao Kojima, Yuko Koyanagi, Kazunori Adachi, Kazuyuki Imamura, and Noriaki Koshikawa. Presynaptic and postsynaptic modulation of glutamatergic synaptic transmission by activation of α1-and β-adrenoceptors in layer v pyramidal neurons of rat cerebral cortex. *Synapse*, 63(4):269–281, 2009.
[37] Masayuki Kobayashi, Kazuyuki Imamura, Tokio Sugai, Norihiko Onoda, Masao Yamamoto, Shoji Komai, and Yasuyoshi Watanabe. Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. *European Journal of Neuroscience*, 12(1):264–272, 2000.
[38] Michael E Hasselmo, Christiane Linster, Madhvi Patil, Daveena Ma, and Milos Cekic. Noradrenergic suppression of synaptic transmission may influence cortical signal-to-noise ratio. *Journal of neurophysiology*, 77(6):3326–3339, 1997.
[39] James FA Poulet, Laura MJ Fernandez, Sylvain Crochet, and Carl CH Petersen. Thalamic control of cortical states. *Nature neuroscience*, 15(3):370–372, 2012.
[40] Dennis B Nestvogel and David A McCormick. Visual thalamocortical mechanisms of waking state-dependent activity and alpha oscillations. *Neuron*, 110(1):120–138, 2022.
[41] Gordon H Petty, Amanda K Kinnischtzke, Y Kate Hong, and Randy M Bruno. Effects of arousal and movement on secondary somatosensory and visual thalamus. *Elife*, 10:e67611, 2021.
[42] Nadia Urbain, Paul A Salin, Paul-Antoine Libourel, Jean-Christophe Comte, Luc J Gentet, and Carl CH Petersen. Whisking-related changes in neuronal firing and membrane potential dynamics in the somatosensory thalamus of awake mice. *Cell reports*, 13(4):647–656, 2015.
[43] Çağatay Aydın, João Couto, Michele Giugliano, Karl Farrow, and Vincent Bonin. Locomotion modulates specific functional cell types in the mouse visual thalamus. *Nature communications*, 9(1):4882, 2018.
[44] Sinem Erisken, Agne Vaiceliunaite, Ovidiu Jurjut, Matilde Fiorini, Steffen Katzner, and Laura Busse. Effects of locomotion extend throughout the mouse early visual system. *Current Biology*, 24(24):2899–2907, 2014.
[45] Benedek Molnár, Péter Sere, Sándor Bordé, Krisztián Koós, Nikolett Zsigri, Péter Horváth, and Magor L Lőrincz. Cell type-specific arousal-dependent modulation of thalamic activity in the lateral geniculate nucleus. *Cerebral cortex communications*, 2(2):tgab020, 2021.
[46] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.
[47] Stefano Panzeri, Monica Moroni, Houman Safaai, and Christopher D Harvey. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience*, 23(9):551–567, 2022.
[48] Brice Bathellier, Lyubov Ushakova, and Simon Rumpel. Discrete neocortical dynamics predict behavioral categorization of sounds. *Neuron*, 76(2):435–449, 2012.
[49] Kenneth D Harris. Cell assemblies of the superficial cortex. *Neuron*, 76(2):263–265, 2012.
[50] Marius Pachitariu, Dmitry R Lyamzin, Maneesh Sahani, and Nicholas A Lesica. State-dependent population coding in primary auditory cortex. *Journal of Neuroscience*, 35(5):2058–2073, 2015.
[51] Kate L Christison-Lagay, Sharath Bennur, and Yale E Cohen. Contribution of spiking activity in the primary auditory cortex to detection in noise. *Journal of neurophysiology*, 118(6):3118–3131, 2017.
[52] Robin AA Ince, Stefano Panzeri, and Christoph Kayser. Neural codes formed by small and temporally precise populations in auditory cortex. *Journal of Neuroscience*, 33(46):18277–18287, 2013.
[53] Edward Zagha and David A McCormick. Neural control of brain state. *Current opinion in neurobiology*, 29:178–186, 2014.
[54] Craig W Berridge and Barry D Waterhouse. The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain research reviews*, 42(1):33–84, 2003.
[55] Cody Slater, Yuxiang Liu, Evan Weiss, Kunpeng Yu, and Qi Wang. The neuromodulatory role of the noradrenergic and cholinergic systems and their interplay in cognitive functions: A focused review. *Brain Sciences*, 12(7):890, 2022.
[56] Kazue Semba. The cholinergic basal forebrain: a critical role in cortical arousal. *The basal forebrain: Anatomy to function*, pages 197–218, 1991.
[57] Susan J Sara. The locus coeruleus and noradrenergic modulation of cognition. *Nature reviews neuroscience*, 10(3): 211–223, 2009.
[58] David A McCormick. Cholinergic and noradrenergic modulation of thalamocortical processing. *Trends in neurosciences*, 12(6):215–221, 1989.
[59] David A McCormick and Thierry Bal. Sleep and arousal: thalamocortical mechanisms. *Annual review of neuroscience*, 20(1):185–215, 1997.
[60] James FA Poulet and Sylvain Crochet. The cortical states of wakefulness. *Frontiers in systems neuroscience*, 12:64, 2019.

[61] Danqing Yang, Chao Ding, Guanxiao Qi, and Dirk Feldmeyer. Cholinergic and adenosinergic modulation of synaptic release. *Neuroscience*, 456:114–130, 2021.
[62] Marina R Picciotto, Michael J Higley, and Yann S Mineur. Acetylcholine as a neuromodulator: cholinergic signaling shapes nervous system function and behavior. *Neuron*, 76(1):116–129, 2012.
[63] Cristina Colangelo, Polina Shichkova, Daniel Keller, Henry Markram, and Srikanth Ramaswamy. Cellular, synaptic and network effects of acetylcholine in the neocortex. *Frontiers in neural circuits*, 13:24, 2019.
[64] John O'Donnell, Douglas Zeppenfeld, Evan McConnell, Salvador Pena, and Maiken Nedergaard. Norepinephrine: a neuromodulator that boosts the function of multiple cell types to optimize cns performance. *Neurochemical research*, 37: 2496–2512, 2012.
[65] Farzan Nadim and Dirk Bucher. Neuromodulation of neurons and synapses. *Current opinion in neurobiology*, 29:48–56, 2014.
[66] Humberto Salgado, Mario Treviño, and Marco Atzori. Layer-and area-specific actions of norepinephrine on cortical synaptic transmission. *Brain research*, 1641:163–176, 2016.
[67] Michael E Hasselmo and James M Bower. Cholinergic suppression specific to intrinsic not afferent fiber synapses in rat piriform (olfactory) cortex. *Journal of neurophysiology*, 67(5):1222–1229, 1992.
[68] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7): 811–819, 2011.
[69] Massimo Mascaro and Daniel J Amit. Effective neural response function for collective population states. *Network: Computation in Neural Systems*, 10(4):351–373, 1999.
[70] Maurizio Mattia, Pierpaolo Pani, Giovanni Mirabella, Stefania Costa, Paolo Del Giudice, and Stefano Ferraina. Heterogeneous attractor cell assemblies for motor planning in premotor cortex. *Journal of Neuroscience*, 33(27):11155–11168, 2013.
[71] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.
[72] Mark M Churchland, Byron M Yu, John P Cunningham, Leo P Sugrue, Marlene R Cohen, Greg S Corrado, William T Newsome, Andrew M Clark, Paymon Hosseini, Benjamin B Scott, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3):369–378, 2010.
[73] Christine F Khoury, Noelle G Fala, and Caroline A Runyan. Arousal and locomotion differently modulate activity of somatostatin neurons across cortex. *Eneuro*, 10(5), 2023.
[74] Lynn KA Sörensen, Sander M Bohté, Heleen A Slagter, and H Steven Scholte. Arousal state affects perceptual decision-making by modulating hierarchical sensory processing in a large-scale visual system model. *PLoS Computational Biology*, 18(4):e1009976, 2022.
[75] Daniela Saderi, Zachary P Schwartz, Charles R Heller, Jacob R Pennington, and Stephen V David. Dissociation of task engagement and arousal effects in auditory cortex and midbrain. *Elife*, 10:e60153, 2021.
[76] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.
[77] Pei-Ann Lin, Samuel K Asinof, Nicholas J Edwards, and Jeffry S Isaacson. Arousal regulates frequency tuning in primary auditory cortex. *Proceedings of the National Academy of Sciences*, 116(50):25304–25310, 2019.
[78] Gideon Rothschild, Israel Nelken, and Adi Mizrahi. Functional organization and population dynamics in the mouse primary auditory cortex. *Nature neuroscience*, 13(3):353–360, 2010.
[79] Akihiro Funamizu, Fred Marbach, and Anthony M Zador. Stable sound decoding despite modulated sound representation in the auditory cortex. *Current Biology*, 33(20):4470–4483, 2023.
[80] Artur Luczak, Peter Barthó, and Kenneth D Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron*, 62(3):413–425, 2009.
[81] Daniel E Winkowski and Patrick O Kanold. Laminar transformation of frequency organization in auditory cortex. *Journal of Neuroscience*, 33(4):1498–1508, 2013.
[82] 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):e68, 2005.
[83] Ho Ko, Sonja B Hofer, Bruno Pichler, Katherine A Buchanan, P Jesper Sjöström, and Thomas D Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011.
[84] Yumiko Yoshimura, Jami LM Dantzker, and Edward M Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433(7028):868–873, 2005.
[85] Lee Cossell, Maria Florencia Iacaruso, Dylan R Muir, Rachael Houlton, Elie N Sader, Ho Ko, Sonja B Hofer, and Thomas D Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518 (7539):399–403, 2015.
[86] 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.
[87] 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.
[88] Jae-eun Kang Miller, Inbal Ayzenshtat, Luis Carrillo-Reid, and Rafael Yuste. Visual stimuli recruit intrinsically generated cortical ensembles. *Proceedings of the National Academy of Sciences*, 111(38):E4053–E4061, 2014.
[89] Jason N MacLean, Brendon O Watson, Gloster B Aaron, and Rafael Yuste. Internal dynamics determine the cortical response to thalamic stimulation. *Neuron*, 48(5):811–823, 2005.

[90] Artur Luczak, Peter Bartho, and Kenneth D Harris. Gating of sensory input by spontaneous cortical activity. *Journal of Neuroscience*, 33(4):1684–1695, 2013.
[91] Shuzo Sakata and Kenneth D Harris. Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron*, 64(3):404–418, 2009.
[92] Luca Mazzucato. Neural mechanisms underlying the temporal organization of naturalistic animal behavior. *Elife*, 11: e76577, 2022.
[93] Liam Lang, Giancarlo La Camera, and Alfredo Fontanini. Temporal progression along discrete coding states during decision-making in the mouse gustatory cortex. *PLOS Computational Biology*, 19(2):e1010865, 2023.
[94] Vahid Rostami, Thomas Rost, Felix Johannes Schmitt, Sacha Jennifer van Albada, Alexa Riehle, and Martin Paul Nawrot. Spiking attractor model of motor cortex explains modulation of neural and behavioral variability by prior target information. *Nature Communications*, 15(1):6304, 2024.
[95] Zachary P Schwartz, Brad N Buran, and Stephen V David. Pupil-associated states modulate excitability but not stimulus selectivity in primary auditory cortex. *Journal of neurophysiology*, 123(1):191–208, 2020.
[96] 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.
[97] Célia Gasselin, Benoît Hohl, Arthur Vernet, Sylvain Crochet, and Carl CH Petersen. Cell-type-specific nicotinic input disinhibits mouse barrel cortex during active sensing. *Neuron*, 109(5):778–787, 2021.
[98] David A McCormick, Zhong Wang, and John Huguenard. Neurotransmitter control of neocortical neuronal activity and excitability. *Cerebral cortex*, 3(5):387–398, 1993.
[99] David A McCormick and David A Prince. Acetylcholine induces burst firing in thalamic reticular neurones by activating a potassium conductance. *Nature*, 319(6052):402–405, 1986.
[100] Jason C Wester and Chris J McBain. Behavioral state-dependent modulation of distinct interneuron subtypes and consequences for circuit function. *Current opinion in neurobiology*, 29:118–125, 2014.
[101] Gabrielle J Gutierrez, Timothy O'Leary, and Eve Marder. Multiple mechanisms switch an electrically coupled, synaptically inhibited neuron between competing rhythmic oscillators. *Neuron*, 77(5):845–858, 2013.
[102] Carina Curto, Shuzo Sakata, Stephan Marguet, Vladimir Itskov, and Kenneth D Harris. A simple model of cortical dynamics explains variability and state dependence of sensory responses in urethane-anesthetized auditory cortex. *Journal of neuroscience*, 29(34):10600–10612, 2009.
[103] David M Schneider, Anders Nelson, and Richard Mooney. A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature*, 513(7517):189–194, 2014.
[104] John M Beggs. The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1864):329–343, 2008.
[105] Jordan O'Byrne and Karim Jerbi. How critical is brain criticality? *Trends in Neurosciences*, 2022.
[106] Woodrow L Shew and Dietmar Plenz. The functional benefits of criticality in the cortex. *The neuroscientist*, 19(1): 88–100, 2013.
[107] Miguel A Munoz. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3): 031001, 2018.
[108] Nils Bertschinger and Thomas Natschläger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural computation*, 16(7):1413–1436, 2004.
[109] LF Abbott. Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Physical Review E*, 84(5):051908, 2011.
[110] Nergis Tomen, David Rotermund, and Udo Ernst. Marginally subcritical dynamics explain enhanced stimulus discriminability under attention. *Frontiers in systems neuroscience*, 8:151, 2014.
[111] Zhuda Yang, Junhao Liang, and Changsong Zhou. Critical avalanches in excitation-inhibition balanced networks reconcile response reliability with sensitivity for optimal neural representation. *Physical Review Letters*, 134(2):028401, 2025.
[112] Mircea Steriade, Angel Nunez, and Florin Amzica. A novel slow (¡ 1 hz) oscillation of neocortical neurons in vivo: depolarizing and hyperpolarizing components. *Journal of neuroscience*, 13(8):3252–3265, 1993.
[113] Mircea Steriade, Igor Timofeev, and François Grenier. Natural waking and sleep states: a view from inside neocortical neurons. *Journal of neurophysiology*, 85(5):1969–1985, 2001.
[114] Artur Luczak, Peter Barthó, Stephan L Marguet, György Buzsáki, and Kenneth D Harris. Sequential structure of neocortical spontaneous activity in vivo. *Proceedings of the National Academy of Sciences*, 104(1):347–352, 2007.
[115] Carl CH Petersen, Thomas TG Hahn, Mayank Mehta, Amiram Grinvald, and Bert Sakmann. Interaction of sensory responses with spontaneous depolarization in layer 2/3 barrel cortex. *Proceedings of the National Academy of Sciences*, 100(23):13638–13643, 2003.
[116] Gabriela Mochol, Ainhoa Hermoso-Mendizabal, Shuzo Sakata, Kenneth D Harris, and Jaime De la Rocha. Stochastic transitions into silence cause noise correlations in cortical circuits. *Proceedings of the National Academy of Sciences*, 112 (11):3529–3534, 2015.
[117] Yan-Liang Shi, Nicholas A Steinmetz, Tirin Moore, Kwabena Boahen, and Tatiana A Engel. Cortical state dynamics and selective attention define the spatial pattern of correlated variability in neocortex. *Nature communications*, 13(1):44, 2022.
[118] Stephan L Marguet and Kenneth D Harris. State-dependent representation of amplitude-modulated noise stimuli in rat auditory cortex. *Journal of Neuroscience*, 31(17):6414–6420, 2011.
[119] Jermyn Z See, Craig A Atencio, Vikaas S Sohal, and Christoph E Schreiner. Coordinated neuronal ensembles in primary auditory cortical columns. *Elife*, 7:e35587, 2018.

[120] Hesam Setareh, Moritz Deger, Carl CH Petersen, and Wulfram Gerstner. Cortical dynamics in presence of assemblies of densely connected weight-hub neurons. *Frontiers in computational neuroscience*, 11:52, 2017.
[121] Garrett T Neske, Dennis Nestvogel, Paul J Steffan, and David A McCormick. Distinct waking states for strong evoked responses in primary visual cortex and optimal visual detection performance. *Journal of Neuroscience*, 39(50):10044–10059, 2019.
[122] Mu Zhou, Feixue Liang, Xiaorui R Xiong, Lu Li, Haifu Li, Zhongju Xiao, Huizhong W Tao, and Li I Zhang. Scaling down of balanced excitation and inhibition by active behavioral states in auditory cortex. *Nature neuroscience*, 17(6): 841–850, 2014.
[123] James Bigelow, Ryan J Morrill, Jefferson Dekloe, and Andrea R Hasenstaub. Movement and vip interneuron activation differentially modulate encoding in mouse auditory cortex. *eNeuro*, 6(5), 2019.
[124] Iryna Yavorska and Michael Wehr. Effects of locomotion in auditory cortex are not mediated by the vip network. *Frontiers in neural circuits*, 15:618881, 2021.
[125] Jacob Reimer, Emmanouil Froudarakis, Cathryn R Cadwell, Dimitri Yatsenko, George H Denfield, and Andreas S Tolias. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *neuron*, 84(2):355–362, 2014.
[126] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.
[127] 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.
[128] Corbett Bennett, Sergio Arroyo, and Shaul Hestrin. Subthreshold mechanisms underlying state-dependent modulation of visual responses. *Neuron*, 80(2):350–357, 2013.
[129] Pierre-Olivier Polack, Jonathan Friedman, and Peyman Golshani. Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience*, 16(9):1331–1339, 2013.
[130] Oliver Rübel, Andrew Tritt, Ryan Ly, Benjamin K. Dichter, Satrajit Ghosh, Lawrence Niu, Pamela Baker, Ivan Soltesz, Lydia Ng, Karel Svoboda, Loren Frank, and Kristofer E. Bouchard. The Neurodata Without Borders ecosystem for neurophysiological data science. *eLife*, 11:e78362, 2022. doi:10.7554/eLife.78362.
[131] JJ Jun, NA Steinmetz, JH Siegle, DJ Denman, M Bauza, B Barbarits, AK Lee, CA Anastassiou, A Andrei, Ç Aydın, M Barbic, TJ Blanche, V Bonin, J Couto, B Dutta, Gratiy SL, DA Gutnisky, M Häusser, B Karsh, P Ledochowitsch, CM Lopez, C Mitelut, S Musa, M Okun, M Pachitariu, PD Putzeys J, Rich, C Rossant, WL Sun, K Svoboda, M Carandini, KD Harris, C Koch, J O'Keefe, and TD Harris. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551:232–236, 2017.
[132] Joshua H Siegle, Aarón Cuevas López, Yogi A Patel, Kirill Abramov, Shay Ohayon, and Jakob Voigts. Open ephys: an open-source, plugin-based platform for multichannel electrophysiology. *Journal of neural engineering*, 14(4):045003, 2017.
[133] Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576:266–273, 2019.
[134] Kip A. Ludwig, Rachel M. Miriani, Nicholas B. Langhals, Michael D. Joseph, David J. Anderson, and Daryl R. Kipke. Using a common average reference to improve cortical neuron recordings from microelectrode array. *Journal of Neurophysiology*, 101(3):1679–1689, 2009.
[135] Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, and Kenneth D Harris. Fast and accurate spike sorting of high-channel count probes with kilosort. *Advances in neural information processing systems*, 29, 2016.
[136] Marius Pachitariu, Shashwat Sridhar, Jacob Pennington, and Carsen Stringer. Spike sorting with kilosort4. *Nature methods*, 21(5):914–921, 2024.
[137] K.B.J. Franklin and G. Paxinos. *The Mouse Brain in Stereotaxic Coordinates*. Academic Press, 1997.
[138] Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185, 2009.
[139] Amelia J Christensen and Jonathan W Pillow. Reduced neural activity but improved coding in rodent higher-order visual cortex during locomotion. *Nature communications*, 13(1):1676, 2022.
[140] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.
[141] Alfonso Renart, Nicolas Brunel, and Xiao-Jing Wang. Mean-field theory of irregularly spiking neuronal populations and working memory in recurrent cortical networks. *Computational neuroscience: A comprehensive approach*, pages 431–490, 2004.
[142] Marina Vegué and Alex Roxin. Firing rate distributions in spiking networks with heterogeneous connectivity. *Physical Review E*, 100(2):022208, 2019.
[143] Nicolas Brunel and Simone Sergi. Firing frequency of leaky intergrate-and-fire neurons with synaptic current dynamics. *Journal of theoretical Biology*, 195(1):87–95, 1998.
[144] Martin P Nawrot, Clemens Boucsein, Victor Rodriguez Molina, Alexa Riehle, Ad Aertsen, and Stefan Rotter. Measurement of variability dynamics in cortical spike trains. *Journal of neuroscience methods*, 169(2):374–390, 2008.
[145] Partha Mitra. *Observed brain dynamics*. Oxford University Press, 2007.
[146] Hemant Bokil, Peter Andrews, Jayant E Kulkarni, Samar Mehta, and Partha P Mitra. Chronux: a platform for analyzing neural signals. *Journal of neuroscience methods*, 192(1):146–151, 2010.
[147] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy. *nature*, 585(7825): 357–362, 2020.

[148] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[149] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[150] Ariel Rokem, M Trumpis, and F Perez. Nitime: time-series analysis for neuroimaging data. In *Proceedings of the 8th Python in Science Conference*, volume 68, page 75, 2009.