bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Circuit-based models of shared variability in cortical networks

Chengcheng Huang[1,2], Douglas A. Ruff[2,3], Ryan Pyle[4], Robert Rosenbaum[4,5], Marlene R. Cohen[2,3] and Brent Doiron[1,2*]

[1]Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA
[2]Center for the Neural Basis of Cognition, Pittsburgh, PA, USA
[3]Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA
[4]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA
[5]Interdisciplinary Center for Network Science and Applications, University of Notre Dame, Notre Dame, IN, USA

*To whom correspondence should be addressed; E-mail: bdoiron@pitt.edu.

**A pervasive yet puzzling feature of cortical circuits is that despite their complex wiring, population-wide shared spiking variability is low dimensional. Neuronal variability is often used as a probe to understand how recurrent circuitry supports network dynamics. However, current models cannot internally produce low dimensional shared variability, and rather assume that it is inherited from outside the circuit. We analyze population recordings from the visual pathway where directed attention differentially modulates shared variability within and between areas, which is difficult to explain with externally imposed variability. We show that if the spatial and temporal scales of inhibitory coupling match physiology, network models capture the low dimensional shared**

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**variability of our population data. Our theory provides a critical link between measured cortical circuit structure and recorded population activity.**

**One Sentence Summary:** Circuit models with spatio-temporal excitatory and inhibitory interactions generate population variability that captures recorded neuronal activity across cognitive states.

# Introduction

The trial-to-trial variability of neuronal responses gives a critical window into how the circuit structure connecting neurons drives brain activity. This idea combined with the widespread use of population recordings has prompted a deep interest in how variability is distributed over a population (*1, 2*). There has been a proliferation of data sets where the shared variability over a population is low dimensional (*3–7*), meaning that neuronal activity waxes and wanes as a group. How cortical networks generate low dimensional shared variability is currently unknown.

Theories of cortical variability can be broadly separated into two categories: ones where variability is internally generated through recurrent network interactions (Fig. 1Ai) and ones where variability originates external to the network (Fig. 1Aii). Networks of spiking neuron models where strong excitation is balanced by opposing recurrent inhibition produce high single neuron variability through internal mechanisms (*8–10*). However, these networks famously enforce an asynchronous solution, and as such fail to explain population-wide shared variability (*11–13*). This lack of success is contrasted with the ease of producing arbitrary correlation structure from external sources. Indeed, many past cortical models assume a global fluctuation from an external source (*2, 7, 14–16*), and accurately capture the structure of population data. However, such phenomenological models are circular, with an assumption of variability from

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

an unobserved source explaining the variability in a recorded population.

Determining whether output variability is internally generated through network interactions or externally imposed upon a network is a difficult problem, where single area population recordings may preclude any definitive solution (Fig. 1Ai vs Aii). In this study we consider attention-mediated shifts in population variability obtained from simultaneous recordings of visual area V1 and the middle temporal area MT (*17*). These data provide novel constraints for how shared variability is distributed within and across populations (Fig. 1Aiii), strongly suggesting that a component is internally generated within an area. We show that networks with spatially dependent coupling (*12, 18*) and synapses with temporal kinetics that match physiology naturally produce low dimensional population-wide variability. Further, attention-mediated top-down modulation of inhibitory neurons (*7*) in our model provides a parsimonious mechanism that controls population-wide variability in agreement with experimental results.

There is a long standing research program aimed at providing a circuit-based understanding for cortical variability (*8–10, 19*). Our work is a critical advance through providing a mechanistic theory for the genesis and propagation of realistic low dimensional population-wide shared variability based on (from) established circuit structure.

# Results

## Externally imposed or internally generated shared variability?

Directed attention reduces the mean spike count correlation coefficient between neuron pairs in visual area V4 during an orientation detection task (Fig. 1Bi; see (*20*)). Further, most of the shared variability across the population is well described by a single latent variable (*4, 7*), consistent with findings from other cortices (*3,5,6*). Thus motivated, we represent the aggregate population response with a scalar random variable $R = X + \beta H$, where $X$ is a stimulus input and $H$ is a hidden source of fluctuations (with strength $\beta$) (Fig. 1Bii). In this simple model

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the trial-to-trial fluctuations are inherited from both $X$ and $H$, but we model attention as only reducing the variance of $H$ (Var$(H)$). There is a large range of parameter values that match the $\sim 30\%$ reduction in Var$(R)$ reported in the V4 data (Fig. 1Biii, blue curve; see Supplemental Text). Certain parameter choices are unreasonable (pink region in Fig. 1Biii), such as $\beta$ being overly large so that $R$ is no longer driven by $X$, or Var$(H) \to 0$ in the attended state, requiring the area that produces $H$ to be silent. Fortunately, there are moderate $\beta$ and Var$(H)$ choices that capture the data (section of the blue curve that is not in the pink region in Fig. 1Biii), meaning that a one-dimensional external latent variable readily explains the data.

Multi-electrode recordings from visual area MT and V1 during an attention modulated task (*17*) also show a reduction of shared variability for neuron pairs within an area (Fig. 1Ci, Top), recapitulating the results observed in V4. However, when V1 and MT neurons are jointly recorded, there is an attention-mediated *increase* of spike count correlations across areas (Fig. 1Ci, Bottom). Returning to the population model with $R$ modeling MT, we augment the model with V1 being the input $X = X_0 + \kappa H$ (Fig. 1Cii). Here $\kappa$ denotes how much the hidden variable is directly shared between areas, and $X_0$ is the variability in $X$ that is independent of $H$. For our linear model we calculate the constraint curves for Var$(R)$ and Cov$(R, X)$ that match the MT-MT and V1-MT data sets (blue curves, Fig. 1Ciii; see Supplementary Text). The constraints require our model to assume both a large influence of $H$ on $R$ and a large attentional modulation of $H$ (pink region in Fig. 1Ciii). This tightening of model assumptions reflects a compromise between an attention-mediated increase in variability transfer from $X \to R$ so that Cov$(R, X)$ increases and a decrease in Var$(H)$ so that Var$(R)$ decreases. This compromise can be mitigated by setting $\kappa$ to be small, meaning that a large component of the fluctuations in $R$ is private from those in $X$ (Fig. 1Ciii).

The source of private variability to an area may still be external to that area, and there are several experimental (*21*) and theoretical (*16*) studies that identify sources of top-down

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

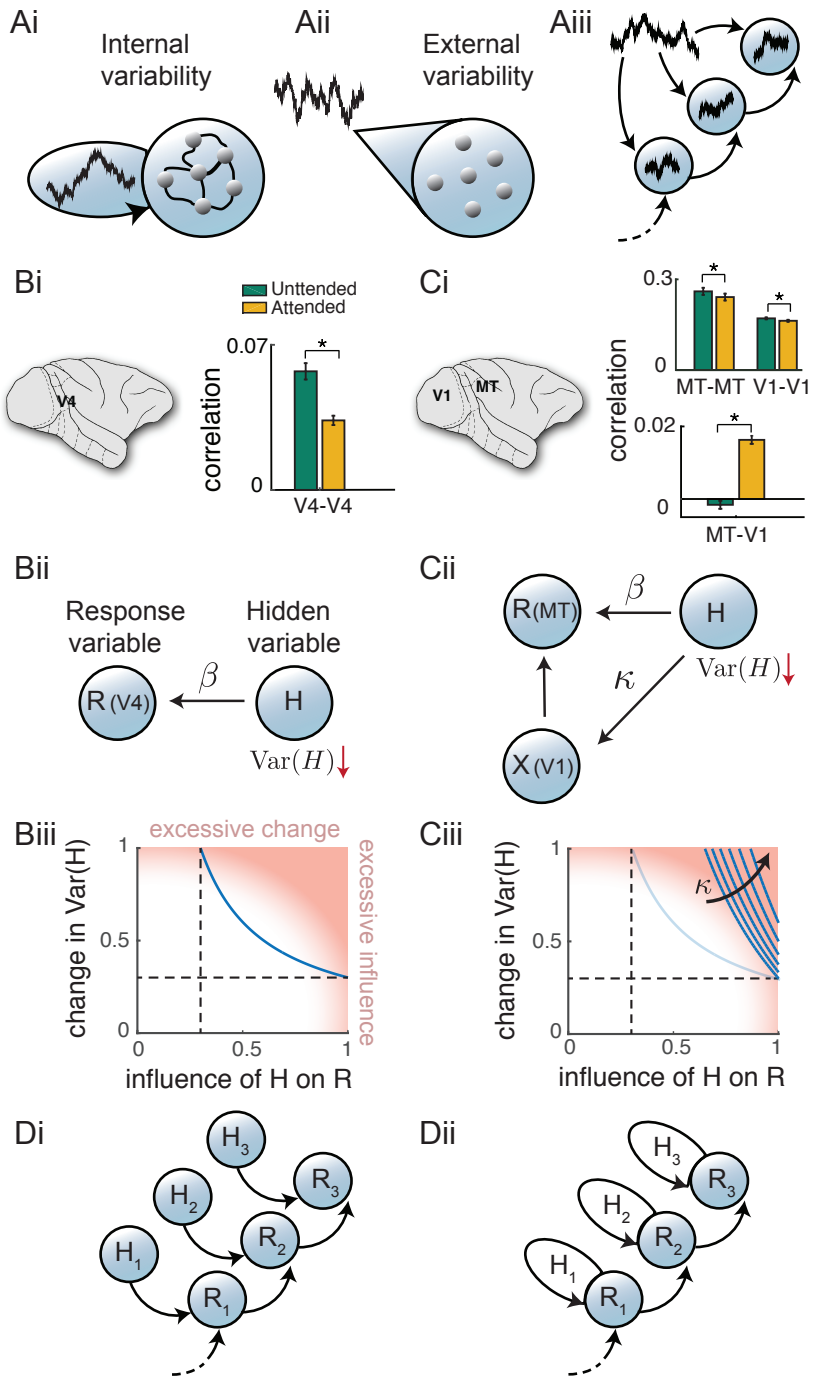


Figure 1: Caption is on next page.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: **Models of shared variability**. **(A)** Variability may either be internally generated within a population **(Ai)** or externally imposed upon a population **(Aii)**. New model constraints emerge by accounting how variability is distributed and modulated across several populations **(Aiii)**. **(Bi)** Attention decreases spike count correlation $r_{\mathrm{SC}}$ obtained from multi-electrode array recording from V4 (n=72765 pairs, Wilcoxon rank-sum test, $p = 3.3 \times 10^{-6}$). **(Bii)** Hidden variable model where the response variability $R$ (modeling V4) comes from a hidden variable $H$ with influence $\beta$. **(Biii)** The attention-mediated reduction in variability in V4 gives a constraint that is a trade off between the reduction in Var$(H)$ and $\beta$ (blue curve). The pink region denotes modulations of $H$ and influence $\beta$ that are excessive. **(Ci)** Top: Mean $r_{\mathrm{SC}}$ between simultaneously recorded MT units (n=270 pairs) and between simultaneously recorded V1 units (n=34404 pairs) in both unattended and attended states (Wilcoxon signed-rank test, MT: $p = 0.017$, V1: $p = 4.9 \times 10^{-6}$). Bottom: Average cross-area $r_{\mathrm{SC}}$ between V1 units and MT units (n=1631 cross-area pairs) in both attentional states (Wilcoxon ranked-sum, $p = 1.4 \times 10^{-4}$). **(Cii)** Hidden variable model for connected areas $X$ (modeling V1) and $R$ (modeling MT); $H$ projects to $X$ with strength $\kappa$. **(Ciii)** The attention mediated changes in $r_{\mathrm{SC}}$ give further constraints on $H$ with the increase in $\kappa$ indicated. Light blue curve is the same as that in Biii for comparison. **(D)** Model of private variability that is externally applied **(Di)** or internally generated **(Dii)** to each area in a feedforward hierarchy. The data in Bi is reproduced from (*20*) and the data in Ci is reproduced from (*17*). Error bars represent the SEM.

variability in cortical circuits. However, if we extend our analysis to a cortical hierarchy then each area requires an external variability 'generator' that is private to that area (Fig. 1Di). A more parsimonious hypothesis is that variability is internally generated within each area (Fig. 1Dii). Below we investigate the circuit mechanics required for low dimensional population-wide shared variability to be an emergent property within a cortical network.

## Population-wide correlations with slow inhibition in spatially ordered networks

Most model cortical networks have disordered connectivity, namely where connection probability is uniform between all neuron pairs (*8, 9, 11, 22*). When such models have a balance between excitation and inhibition they produce spike trains that are irregular in time (Fig. 2Ai), with a broad distribution of firing rates (Fig. 2B, top purple curve), and uncorrelated between neurons (Fig. 2Ai and C, top purple curve). However, there is abundant evidence that cortical connec-

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tivity is spatially ordered with a connection probability falling off with the distance between neuron pairs (*23, 24*). Recently we have extended the theory of balanced networks to include such spatially dependent connectivity (*12, 18*). Briefly, we model a two dimensional array of integrate-and-fire style neurons that receive both feedforward projections from a layer of external Poisson processes as well as recurrent projections within the network (see Methods for a full model description); connection probability of all projections decays like a Gaussian with distance (Fig. 2Aii, left). If the spatial scale of feedforward inputs is narrower than the scale of recurrent projections the asynchronous state no longer exists (*12*), giving way to a solution with weak but spatially structured correlations (Fig. 2Aii and supplemental movie S1). Nevertheless, the mean correlation across all neuron pairs vanishes for large network size (Fig. 2C, bottom purple curve), in stark disagreement with a vast majority of experimental studies (*1, 2*), including the data from V4 (Fig. 1Bi) and MT (Fig. 1Ci).

Many cortical network models assume that the kinetics of inhibitory conductances are *faster* than those of excitatory conductances (*8–11, 25*), including our past models of spatially ordered networks (Fig. 2Aii and (*12*)). However, this assumption is at odds with physiology where excitatory $\alpha$-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPA) have faster kinetics (*26, 27*) than those of the inhibitory $\gamma$-Aminobutyric acid receptors (GABAa) (*28, 29*). In networks with disordered connectivity, when the timescales of excitation and inhibition match experimental values then activity becomes pathologic, with homogeneous firing rates (Fig. 2B, top green curve) and excessive synchrony (Fig. 2Aiii and C, top green curve). This consequence is likely the justification for the faster inhibitory kinetics in model networks.

When a spatially ordered model has synaptic kinetics that respect physiology they produce dynamics that are quite distinct from those of disordered networks. Firing rates are broad (Fig. 2B, bottom green curve) and pairwise correlations are reasonable in magnitude (Fig. 2C, bottom green curve). Further, population-wide turbulent dynamics (Fig. 2Aiv and Supplemental movie

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

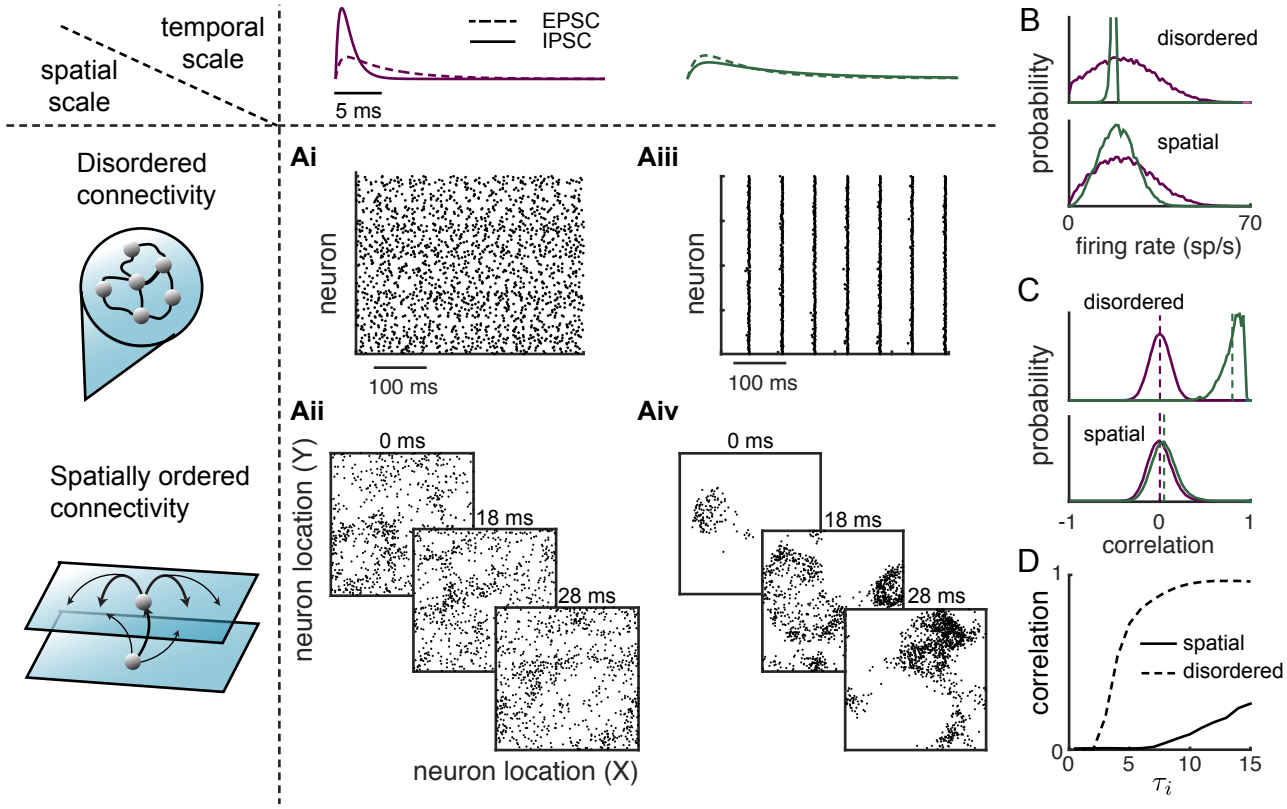


Figure 2: **The spatial and temporal scales of synaptic wiring determine internally generated variability.** Networks of excitatory and inhibitory neuron models were simulated with either disordered connectivity (**Ai** and **Aiii**)) or spatially ordered connectivity (**Aii** and **Aiv**). Networks with fast inhibition ($\tau_i = 1$ ms; **Ai** and **Aii**) were compared to networks with slow inhibition ($\tau_i = 8$ ms; **Aiii** and **Aiv**); in all models the timescale of excitation was $\tau_e = 5$ ms. For spatially disordered models spike train rasters are plotted where no particular neuron ordering is used (**Ai** and **Aiii**). For spatially ordered networks three consecutive spike time raster snapshots are shown with a black dot indicating that the neuron at spatial position $(x, y)$ fired within one millisecond of the time stamp. **(B)** Distributions of firing rates of excitatory neurons in the disordered (top) and spatially ordered (bottom) models, with faster inhibitory kinetics (purple) compared to slower inhibitory kinetics (green). **(C)** Same as B for the distributions of pairwise correlations among the excitatory population. **(D)** Mean correlation among the excitatory population as a function of the inhibitory decay time constant ($\tau_i$).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S2) accompany a small, but nonzero, mean pairwise spike count correlation across the population ($r_{\mathrm{SC}} = 0.04$, Fig. 2C). Indeed, as the timescale of inhibition grows disordered networks show a rapid change in mean pairwise correlation while two dimensional spatially ordered networks show a much more gradual rise in correlation (Fig. 2D). We remark that networks constrained to one spatial dimension also produce excessive synchrony (Fig. S1), meaning that two (or more) spatial dimensions are required for robustly low but nonzero correlations. In sum, when realistic spatial synaptic connectivity is paired with realistic temporal synaptic kinetics in spatially ordered networks, internally generated population dynamics produces spiking dynamics whose marginal and pairwise variability conform to experimental results.

## Low dimensional variability and attentional modulation

The V1 and MT network is modeled by extending the spatially ordered balanced networks with slow inhibition to include three layers: a bottom layer of independent Poisson processes modeling thalamus, and middle and top layers of integrate and fire style neurons modeling V1 and MT, respectively (Fig. 3A and see Methods). We follow our past work with simplified firing rate networks (*7*) and model a top-down attentional signal as an overall depolarization to inhibitory neurons in the MT layer (Fig. 3A). This mimics cholinergic pathways that primarily affect interneurons (*30*) and are thought to be engaged during attention (*7*). The increased recruitment of inhibition during attention reduces the population-wide fluctuations in the MT layer (Fig. 3B) and the pairwise spike count correlation of MT-MT neuron pairs (Fig. 3C), while it increases the correlation of V1-MT neuron pairs (Fig. 3D), thereby capturing the main aspects of the V1-MT dataset (Fig. 1Ci).

To this point we have measured shared variability with only population averaged pairwise spike count correlation, as done in many other studies (*1, 2*). A deeper understanding of how shared variability is distributed over the population comes from dimensionality reduction

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

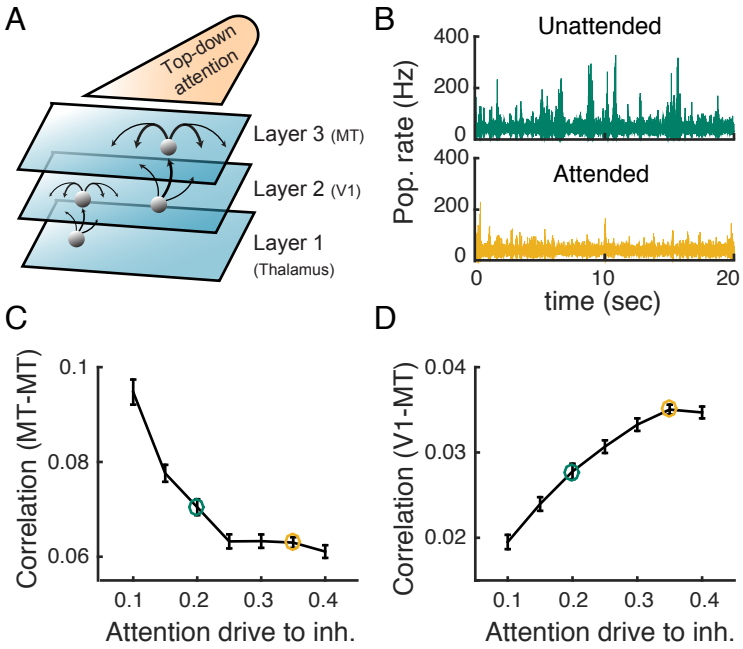


Figure 3: **Top-down depolarization of MT inhibitory neurons capture the differential attentional modulation of shared variability within and across V1 and MT**. **(A)** Thalamus, V1, and MT are modeled in a three layer hierarchy of spatially ordered balanced networks. Top-down attentional modulation is modeled as a depolarization to MT inhibitory neurons ($\mu_I$). In both V1 and MT the recurrent projections are broader than feedforward projections and recurrent inhibition is slower than excitation. **(B)** Population averaged firing rate fluctuations from MT in the unattended state ($\mu_I = 0.2$, green) and the attended state ($\mu_I = 0.35$, orange). **(C)** Mean spike count correlation ($r_{\mathrm{SC}}$) of excitatory neuron pairs in MT decreases with attentional modulation. **(D)** Mean $r_{\mathrm{SC}}$ between the excitatory neurons in MT and the excitatory neurons in V1 increases with attention. Error bars are standard error.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tools (*31*). We partition the covariance matrix into the shared variability among the population and the private noise to each neuron; the eigenvalues of the shared covariance matrix represent the variance along each dimension (or latent variable), while the corresponding eigenvectors represent the projection weights of the latent variables onto each neuron. Applying these techniques to the multi-electrode V4 data (*20*) shows a single dominant eigenmode (Fig. 4Ai, results from each session in Fig. S2). This mode influences most of the neurons in the population in the same way (Fig. 4Aii, weights are dominant positive), and after subtracting the first mode the mean residual covariances are very small (Fig. 4Aiii). Finally, attention affects population variability primarily by quenching this dominant mode (Fig. 4Ai, compare orange and green curves).

The dimensionality of shared variability offers a strong test to our cortical model. We analyzed the spike count covariance matrix from a subsampling of the spike trains from the third layer of our network model ($n = 50$ neurons). The network with slow inhibition produced shared variability with a clear dominant eigenmode that mimicked many of the core features observed in the V4 data (Fig. 4Bi-iii). Further, the top-down attentional modulation of inhibition also suppressed this dominant mode (Fig. 4Bi, compare orange and green curves). This agreement between model and data broke down when either inhibition was faster than the excitation (Fig. 4Ci-iii), or inhibitory projections in the third layer were spatially broader than those of excitation (Fig. 4Di-iii).

To gain intuition about the mechanics behind low dimensional shared variability we considered a two-dimensional firing rate model that captured the main dynamics of the spiking network but was amenable to modern techniques in dynamical system theory (see Supplementary Text). In particular, we considered how the stability of the asynchronous state is lost as the temporal and spatial scales of inhibition, $\tau_i$ and $\sigma_i$, are varied. When $\tau_i$ and $\sigma_i$ match that of excitation, a stable firing rate solution exists (Fig. 4E, grey region). When either $\tau_i$ or $\sigma_i$

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

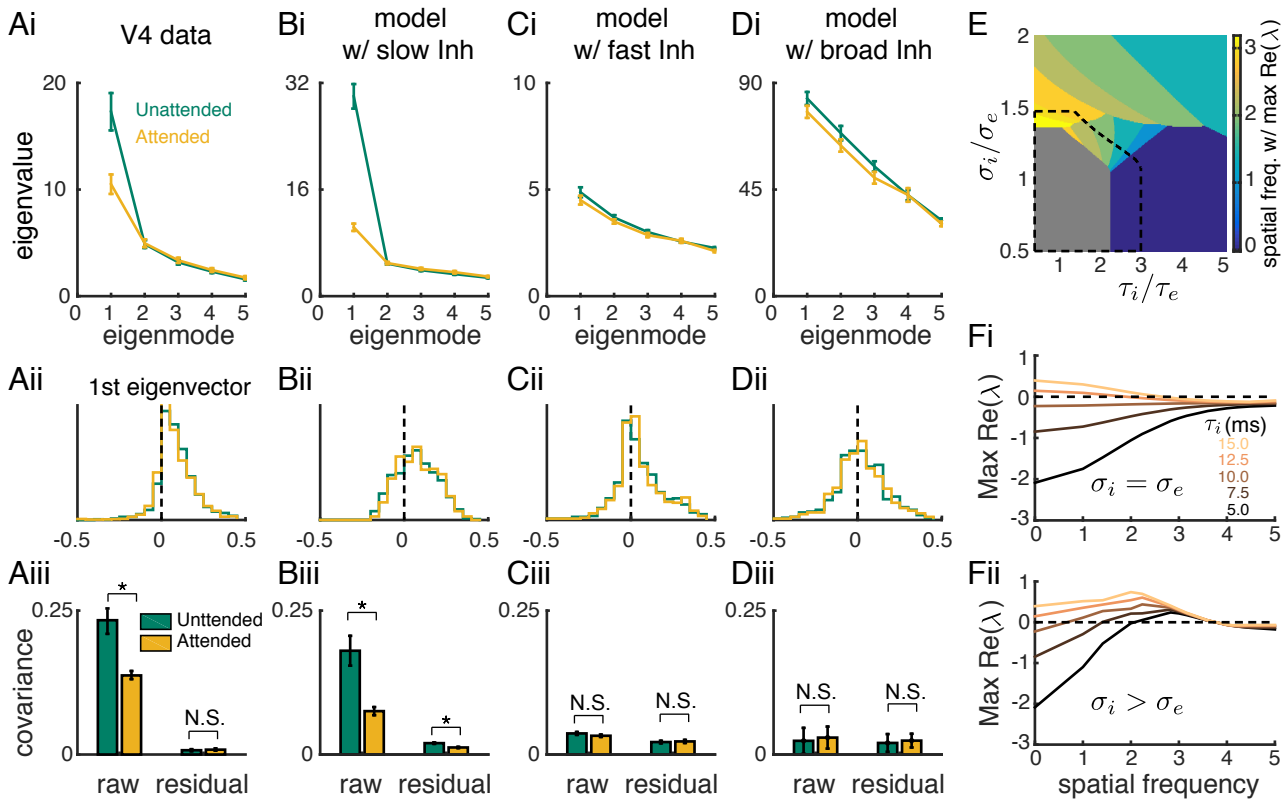


Figure 4: **Internally generated shared variability from the model network is low-dimensional.** **(Ai)** Eigenvalues of the first five modes of the shared component of the spike count covariance matrix from the V4 data (*20*). Both unattended (green) and attended (orange) data sets are analyzed; the population had $n = 43 \pm 15$ neurons. Error bars are standard error. **(Aii)** The vector elements for the first (dominant) eigenmode. **(Aiii)** The mean covariance in attended and unattended states before (raw) and after (residual) subtracting the first latent variable. **(B-D)** Same as A but for the three layer model with slow inhibition **(Bi-iii)**, model with fast inhibition **(Ci-iii)** and model with slow and broad inhibition **(Di-iii)**. Wilcoxon rank-sum test is applied on comparing data from the attended and unattended states: mean covariance: $p = 0.0013$ (Aiii), $p = 1.78 \times 10^{-22}$ (Biii), $p = 0.7798$ (Ciii) and $p = 0.5850$ (Diii); residual: $p = 0.7477$ (Aiii), $p = 5.40 \times 10^{-4}$ (Biii), $p = 0.8796$ (Ciii) and $p = 0.5326$ (Diii). **(E)** Bifurcation diagram of a firing rate model as a function of the inhibitory decay time scale $\tau_i$ and inhibitory projection width $\sigma_i$. The excitatory projection width and time constants are fixed at $\sigma_e = 0.1$ and $\tau_e = 5$ ms, respectively. Color represents the spatial frequency with the largest real part of eigenvalue and gray region is stable. Top-down modulation of inhibitory neurons modeling attention expands the stable region (black dashed). **(Fi)** The real part of eigenvalues as a function of spatial frequency for increasing $\tau_i$ when $\sigma_i$ equals $\sigma_e$. **(Fii)** Same as Fi for $\sigma_i$ larger than $\sigma_e$.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

increase, this stability is first lost at a particular spatial frequency (Fig. 4E, colored regions). When $\tau_i$ increases and excitation and inhibition project similarly ($\sigma_i = \sigma_e$) then firing rate stability is lost first at zero spatial frequency (Fig. 4Fi). This creates population dynamics with a broad spatial pattern, allowing variability to be shared over the entire population, in agreement with data. In contrast, when $\tau_i$ increases and inhibition is lateral ($\sigma_i > \sigma_e$) then stability is lost at a characteristic spatial scale (Fig. 4Fii). Finally, attentional modulations that depolarize the inhibitory population expands the stable region in the bifurcation diagram (Fig. 4E, dashed black line). In other words, attention increases the domain of firing rate stability (*7*), quenching any tendency for patterned firing rate dynamics.

## Chaotic population-wide dynamics reflects internally generated variability

The V1 and MT population data lead us to propose that shared variability has a sizable internally generated component (Fig. 1Di, Dii). Further, our analysis of the simplified scalar model showed that attention must quench a large portion of shared variability (Var$(H)$; Fig. 1Ciii), suggestive of a strong, network-wide nonlinearity. To understand how the three layer network model creates variability with these properties we aimed to isolate the sources of externally and internally generated fluctuations from one another. To this end, we fixed the spike train realizations from the first layer (thalamic) neurons as well as the membrane potential states of the second layer (V1), and only the initial membrane potentials of the third layer (MT) neurons were randomized across trials (Fig. 5A). This produced deterministic network dynamics when conditioned on activity from the first two layers, and consequently any trial-to-trial variability is due to mechanics internal to the third layer.

Our three layer model follows classic work in balanced networks (*8, 18, 22, 32*), with spike trains from third layer neurons in both the unattended and attended states having significant trial-to-trial variability despite the frozen layer one and two inputs (Fig. 5B). To investigate how this

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

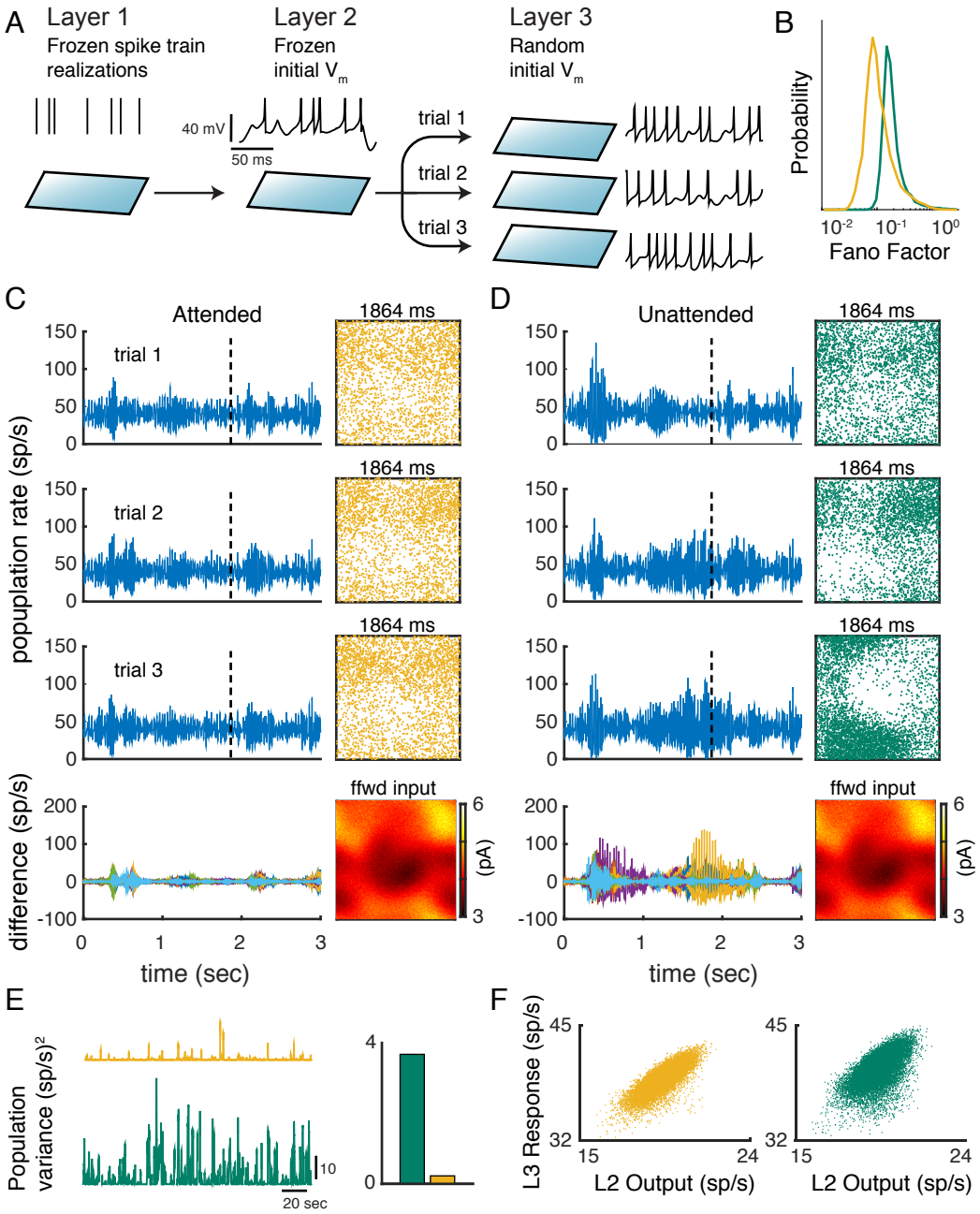


Figure 5: Caption is on next page.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5: **Chaotic population firing rate dynamics is quenched by attention.** **(A)** Schematic of the numerical experiment. The spike train realizations in layer one and the initial states of the membrane potential of layer two neurons are identical across trials, while in each trial we randomized the initial states of the layer three neuron's membrane potentials. **(B)** Fano factor distribution of layer three excitatory neurons in the attended (mean=0.25) and unattended states (mean=0.37) with frozen layer one and two inputs. **(C)** Three representative trials of the layer three excitatory population rates in the attended state (left row 1-3). Bottom row: difference of the population rates across 20 trials. Right (row 1-3): Snapshots of the neuron activity at time point 1864 ms. Each dot is a spike within 2 ms window from the neuron at that location. Right bottom: the synaptic current each layer three neuron receives from layer two at time 1864 ms. **(D)** same as panel C for the network in the unattended state. (E) Trial-to-trial variance of layer three population rates as a function of time; right: mean variance across time. **(F)** The layer three population rate tracks the layer two population rate better in the attended state. Both outputs and responses are smoothed with a 200 ms window.

microscopic (single neuron) variability possibly manifests as macroscopic population activity we considered the trial-to-trial variability of the population-averaged instantaneous firing rate. While the population firing rate is dynamically variable in the attended state, it is nevertheless nearly identical across trials (Fig. 5C, left), resulting in very small trial-to-trial variance of the population rate (Fig. 5E, orange). A reflection of this low population variability is the faithful tracking of the spatiotemporal structure of layer two outputs by layer three responses (Fig. 5C, right; F orange). This tracking supports the high correlation between layer two and three spiking (Fig. 3D).

In contrast, in the unattended state the asynchronous solution is unstable, resulting in population-wide recruited activity. These periods of spatial coherence across the network are trial-to-trial unpredictable, contributing sizable trial-to-trial variability to population activity (Fig. 5D, left; E, green). This internal variability degrades the tracking of layer two outputs (Fig. 5D, right; F, green), ultimately lowering the correlation between layer two and three spiking (Fig. 3D). Thus, while the network model is chaotic in both attended and unattended states, the chaos is population-wide only in the inhibition deprived unattended state.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Discussion

The study of spatiotemporal pattern formation in neuronal populations has a long and rich history of study in both theoretical (*33*) and experimental contexts (*34*). However, a majority of these studies have focused on trial-averaged activity, with tacit assumptions about how spiking variability emerges (but see (*35*) and (*12*)). There is a parallel research program aimed at understanding how variability is an emergent property of recurrent networks (*8, 9, 22, 32*), yet analysis is often restricted to simple networks with disordered connectivity. In this study we have combined these traditions with the goal of understanding the genesis and modulation of population-wide, internally generated cortical variability.

Our solution was to extend classical work in balanced cortical networks (*8, 11*) to include two well accepted truths. First, cortical connectivity has a wiring rule that depends upon neuron pair distance (*23, 24*). Theoretical studies that model distance dependent coupling commonly assume that inhibition projects more broadly than excitation (*33, 35, 36*) (but see (*25*)). However, measurements of local cortical circuitry show that excitation and inhibition project on similar spatial scales (*23, 37*), and long-range excitation is known to project more broadly than inhibition (*38*). Our work shows that this architecture is required for internally generated population variability to be low dimensional (Fig. 4Bi and Di). The second truth is that inhibition has temporal kinetics that are slower than excitation (*26–29*). Past theoretical models of recurrent cortical circuits have assumed that inhibition is not slower than excitation (*8, 11, 25, 39*), including past work from our group (*10, 12*). Consequently, these studies could only capture the residual correlation structure of population recordings once the dominant eigenmode was subtracted (*12, 13*); in these cases the residual accounted for less than ten percent of the true shared variability.

The above narrative is somewhat revisionist; there are several well known theoretical studies

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in disordered networks where one dimensional population-wide correlations do emerge, notably in networks where rhythmic (*40*) or 'up-down' (*36, 39*) dynamics are prominent. However, the lack of wiring structure in such model networks ensures that all neuron pairs experience the same fluctuation, justifying a simple one dimensional mean field. In other words, any shared variability would be one dimensional by construction. Indeed, balanced networks with clustered wiring (*10*) (where simple one dimensional field theories do not apply) show higher dimensional shared variability (*13*).

Rich spatiotemporal chaos is a hallmark feature of systems that are far from equilibrium in fluid mechanics, chemistry and biology (*41*). Networks of model neurons with strong recurrent excitation and inhibition also show chaotic dynamics, though past work has only studied dynamics restricted to the asynchronous state in spatially disordered networks (*22, 42*). Our extension to spatial networks shows that network-wide chaotic activity can be prominent in the weakly correlated state as well. Modern theories of computing in distributed systems often require chaotic dynamics to provide a rich repertoire of network states both in silico (*43*) as well as recently in analogue neuronal-inspired hardware (*44*). Our model provides a framework where the spatial scale of chaotic dynamics is tunable; uncovering the consequences for computation in this framework is an intriguing next step.

# References

1. M. Cohen, A. Kohn, *Nature neuroscience* **14**, 811 (2011).

2. B. Doiron, A. Litwin-Kumar, R. Rosenbaum, G. Ocker, K. Josic, *Nature neuroscience* **19**, 383 (2016).

3. I.-C. Lin, M. Okun, M. Carandini, K. D. Harris, *Neuron* **87**, 644 (2015).

4. N. C. Rabinowitz, R. L. Goris, M. Cohen, E. Simoncelli, *eLife* p. e08998 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. A. S. Ecker, *et al.*, *Neuron* **82**, 235 (2014).

6. M. Okun, *et al.*, *Nature* **521**, 511 (2015).

7. T. Kanashiro, G. K. Ocker, M. R. Cohen, B. Doiron, *eLife* **6** (2017).

8. C. van Vreeswijk, H. Sompolinsky, *Science* **274**, 1724 (1996).

9. D. J. Amit, N. Brunel, *Cerebral cortex* **7**, 237 (1997).

10. A. Litwin-Kumar, B. Doiron, *Nature neuroscience* **15**, 1498 (2012).

11. A. Renart, *et al.*, *Science* **327**, 587 (2010).

12. R. Rosenbaum, M. Smith, A. Kohn, J. Rubin, B. Doiron, *Nature Neuroscience* (2016).

13. R. C. Williamson, *et al.*, *PLOS Computational Biology* **12**, e1005141 (2016).

14. G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, K. D. Miller, *bioRxiv* p. 094334 (2016).

15. A. Ponce-Alvarez, A. Thiele, T. D. Albright, G. R. Stoner, G. Deco, *Proceedings of the National Academy of Sciences* **110**, 13162 (2013).

16. K. Wimmer, *et al.*, *Nature communications* **6** (2015).

17. D. A. Ruff, M. R. Cohen, *Journal of Neuroscience* **36**, 7523 (2016).

18. R. Rosenbaum, B. Doiron, *Physical Review X* **4**, 021039 (2014).

19. M. N. Shadlen, W. T. Newsome, *The Journal of neuroscience* **18**, 3870 (1998).

20. M. Cohen, J. Maunsell, *Nature neuroscience* **12**, 1594 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. C. Gómez-Laberge, A. Smolyanskaya, J. J. Nassi, G. Kreiman, R. T. Born, *Neuron* **91**, 540 (2016).

22. M. Monteforte, F. Wolf, *Physical Review X* **2**, 041007 (2012).

23. R. B. Levy, A. D. Reyes, *Journal of Neuroscience* **32**, 5609 (2012).

24. S. Horvat, *et al.*, *PLoS biology* **14**, e1002512 (2016).

25. S. Lim, M. S. Goldman, *Journal of Neuroscience* **34**, 6790 (2014).

26. J. R. Geiger, J. Lübke, A. Roth, M. Frotscher, P. Jonas, *Neuron* **18**, 1009 (1997).

27. M. C. Angulo, J. Rossier, E. Audinat, *Journal of Neurophysiology* **82**, 1295 (1999).

28. Z. Xiang, J. R. Huguenard, D. A. Prince, *The Journal of Physiology* **506**, 715 (1998).

29. P. A. Salin, D. A. Prince, *Journal of neurophysiology* **75**, 1573 (1996).

30. K. V. Kuchibhotla, *et al.*, *Nature neuroscience* **20**, 62 (2017).

31. J. P. Cunningham, M. Y. Byron, *Nature neuroscience* **17**, 1500 (2014).

32. M. London, A. Roth, L. Beeren, M. Häusser, P. E. Latham, *Nature* **466**, 123 (2010).

33. B. Ermentrout, *Reports on progress in physics* **61**, 353 (1998).

34. T. K. Sato, I. Nauhaus, M. Carandini, *Neuron* **75**, 218 (2012).

35. A. Keane, P. Gong, *Journal of Neuroscience* **35**, 1591 (2015).

36. A. Compte, M. V. Sanchez-Vives, D. A. McCormick, X.-J. Wang, *Journal of neurophysiology* **89**, 2707 (2003).

37. J. Mariño, *et al.*, *Nature neuroscience* **8**, 194 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. W. H. Bosking, Y. Zhang, B. Schofield, D. Fitzpatrick, *Journal of neuroscience* **17**, 2112 (1997).

39. C. Stringer, *et al.*, *Elife* **5**, e19695 (2016).

40. N. Brunel, *Journal of computational neuroscience* **8**, 183 (2000).

41. M. C. Cross, P. C. Hohenberg, *Reviews of modern physics* **65**, 851 (1993).

42. J. Kadmon, H. Sompolinsky, *Physical Review X* **5**, 041030 (2015).

43. D. Sussillo, L. F. Abbott, *Neuron* **63**, 544 (2009).

44. S. Kumar, J. P. Strachan, R. S. Williams, *Nature* **548**, 318 (2017).

# Acknowledgments

NIH Grants CRCNS R01DC015139-01ZRG1 (B.D.), 4R00EY020844- 03 (M.R.C.), R01 EY022930 (M.R.C.), 5T32NS7391-14 (D.A.R.), and Core Grant P30 EY008098; NSF Grants DMS-1517828 (R.R.) and DMS-1517082 (B.D.); Vannevar Bush faculty fellowship N00014-18-1-2002 (B.D.), a Whitehall Foundation Grant (M.R.C.); a Klingenstein-Simons Fellowship (M.R.C.); grants from the Simons Foundation (B.D. and M.R.C.); a Sloan Research Fellowship (M.R.C.); a McKnight Scholar Award (M.R.C.).

# Supplementary materials

Materials and Methods

Supplementary Text

Figs. S1-S2

Movies S1-S2

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Supplementary materials

**This PDF file includes:**

Materials and Methods
Supplementary Text
Figs. S1-S2
Captions for Movies S1-S2

**Other Supplementary Materials for this manuscript includes the following:**

Movies S1-S2

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Materials and Methods

**Network model description.** The network consists of three layers. Layer 1 is modeled by a population of $N_1 = 2500$ excitatory neurons, the spikes of which are taken as independent Poisson processes with a uniform rate $r_1 = 10$ Hz. Layer 2 and Layer 3 are recurrently coupled networks with excitatory ($\alpha = \mathrm{e}$) and inhibitory ($\alpha = \mathrm{i}$) populations of $N_\mathrm{e} = 40000$ and $N_\mathrm{i} = 10000$ neurons, respectively. Each neuron is modeled as an exponential integrate-and-fire (EIF) neuron whose membrane potential is described by:

$$C_m \frac{\mathrm{d}V_j^\alpha}{\mathrm{d}t} = -g_L \left(V_j^\alpha - E_L\right) + g_L \Delta_T e^{\left(V_j^\alpha - V_T\right)/\Delta_T} + I_j^\alpha(t). \tag{1}$$

Each time $V_j^\alpha(t)$ exceeds a threshold $V_\mathrm{th}$, the neuron spikes and the membrane potential is held for a refractory period $\tau_\mathrm{ref}$ then reset to a fixed value $V_\mathrm{re}$. Neuron parameters for excitatory neurons are $\tau_m = C_m/g_L = 15$ ms, $E_L = -60$ mV, $V_T = -50$ mV, $V_\mathrm{th} = -10$ mV, $\Delta_T = 2$ mV, $V_\mathrm{re} = -65$ mV and $\tau_\mathrm{ref} = 1.5$ ms. Inhibitory neurons are the same except $\tau_m = 10$ ms, $\Delta_T = 0.5$ mV and $\tau_\mathrm{ref} = 0.5$ ms. The total current to each neuron is:

$$\frac{I_j^\alpha(t)}{C_m} = \sum_{k=1}^{N_F} \frac{J_{jk}^{\alpha F}}{\sqrt{N}} \sum_n \eta_F\left(t - t_n^{F,k}\right) + \sum_{\beta=\mathrm{e,i}} \sum_{k=1}^{N_\beta} \frac{J_{jk}^{\alpha\beta}}{\sqrt{N}} \sum_n \eta_\beta\left(t - t_n^{\beta,k}\right) + \mu_\alpha, \tag{2}$$

where $N = N_e + N_i$ is the total number of the network population. Postsynaptic current is

$$\eta_\beta(t) = \frac{1}{\tau_{\beta d} - \tau_{\beta r}} \begin{cases} e^{-t/\tau_{\beta\mathrm{d}}} - e^{-t/\tau_{\beta\mathrm{r}}}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{3}$$

where $\tau_\mathrm{er} = 1$ ms, $\tau_\mathrm{ed} = 5$ ms and $\tau_\mathrm{ir} = 1$ ms, $\tau_\mathrm{id} = 8$ ms. The feedforward synapses from Layer 1 to Layer 2 have the same kinetics as the recurrent excitatory synapse, i.e. $\eta_F^{(2)}(t) = \eta_\mathrm{e}(t)$. The feedforward synapses from Layer 2 to Layer 3 have a fast and a slow component.

$$\eta_F^{(3)}(t) = p_\mathrm{f}\eta_\mathrm{e}(t) + p_\mathrm{s}\eta_s(t)$$

with $p_\mathrm{f} = 0.2$, $p_\mathrm{s} = 0.8$. $\eta_s(t)$ has the same form as Eq. 3 with a rise time constant $\tau_r^s = 2$ ms and a decay time constant $\tau_d^s = 100$ ms. The excitatory and inhibitory neurons in Layer 3 receive static current $\mu_e$ and $\mu_i$, respectively.

Neurons on the three layers are arranged on a uniform grid covering a unit square $\Gamma = [0,1] \times [0,1]$. The probability that two neurons, with coordinates $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ respectively, are connected depends on their distance measured periodically on $\Gamma$:

$$p_{\alpha\beta}(\mathbf{x}, \mathbf{y}) = \bar{p}_{\alpha\beta} g(x_1 - y_1; \alpha_\beta) g(x_2 - y_2; \alpha_\beta). \tag{4}$$

Here $\bar{p}_{\alpha\beta}$ is the mean connection probability and

$$g(x; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{k=-\infty}^{\infty} e^{-(x+k)^2/(2\sigma^2)} \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is a wrapped Gaussian distribution. Excitatory and inhibitory recurrent connection widths of Layer 2 are $\alpha_{\text{rec}}^{(2)} := \alpha_{\text{e}}^{(2)} = \alpha_{\text{i}}^{(2)} = 0.1$ and feedforward connection width from Layer 1 to Layer 2 is $\alpha_{\text{ffwd}}^{(2)} = 0.05$. The recurrent connection width of Layer 3 is $\alpha_{\text{rec}}^{(3)} = 0.2$ and the feedforward connection width from Layer 2 to Layer 3 is $\alpha_{\text{ffwd}}^{(3)} = 0.1$. A presynaptic neuron is allowed to make more than one synaptic connection to a single postsynaptic neuron.

The recurrent connectivity of Layer 2 and Layer 3 have the same synaptic strengths and mean connection probabilities. The recurrent synaptic weights are $J_{\text{ee}} = 80$ mV, $J_{\text{ei}} = -240$ mV, $J_{\text{ie}} = 40$ mV and $J_{\text{ii}} = -300$ mV. Recall that individual synapses are scaled with $1/\sqrt{N}$ (Eq. 2); so that, for instance, $J_{ee}/\sqrt{N} \approx 0.36$ mV. The mean connection probabilities are $\bar{p}_{\text{ee}} = 0.01$, $\bar{p}_{\text{ei}} = 0.04$, $\bar{p}_{ie} = 0.03$, $\bar{p}_{\text{ii}} = 0.04$. The out-degrees are $K_{\text{ee}}^{\text{out}} = 400$, $K_{\text{ei}}^{\text{out}} = 1600$, $K_{\text{ie}}^{\text{out}} = 300$ and $K_{\text{ii}}^{\text{out}} = 400$. The feedforward connection strengths from Layer 1 to Layer 2 are $J_{\text{eF}}^{(2)} = 140$ mV and $J_{\text{iF}}^{(2)} = 100$ mV with probabilities $\bar{p}_{\text{eF}}^{(2)} = 0.1$ and $\bar{p}_{\text{iF}}^{(2)} = 0.05$ (out-degrees $K_{\text{eF2}}^{\text{out}} = 4000$ and $K_{\text{eF2}}^{\text{out}} = 500$). The feedforward connection strengths from Layer 2 to Layer 3 are $J_{\text{eF}}^{3} = 25$ mV and $J_{\text{iF}}^{3} = 15$ mV with mean probabilities $\bar{p}_{\text{eF}}^{(3)} = 0.05$ and $\bar{p}_{\text{iF}}^{(3)} = 0.05$ (out-degrees are $K_{\text{eF3}}^{\text{out}} = 2000$ and $K_{\text{iF3}}^{\text{out}} = 500$). Only the excitatory neurons in Layer 2 project to Layer 3.

The spatial model in Fig. 2 contains only Layer 1 and Layer 2. In the model with disordered connectivity, the connection probability between a pair of neurons is $\bar{p}_{\alpha\beta}$, independent of distance. Other parameters are the same as the spatial model. The decay time constant of IPSC ($\tau_{\text{id}}$) was varied from 1 to 15 ms (Fig. 2D). The rise time constant of IPSC ($\tau_{\text{ir}}$) is 1 ms when $\tau_{\text{id}} > 1$ ms and 0.5 ms when $\tau_{\text{id}} = 1$ ms.

The parameters used in Fig. 3C,D are $\mu_i = [0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4]$ pA and $\mu_E = 0$ pA. The mean firing rates in Layer 2 are $r_{\text{e}}^{(2)} = 19$ Hz and $r_{\text{i}}^{(2)} = 9$ Hz. In the further analysis (Fig. 4Bi-Biii and Fig. 5), we used $\mu_I = 0.2$ pA for the unattended state and $\mu_I = 0.35$ pA for the attended state. In simulations of the spatial model with fast inhibition (Fig. 4Ci-Ciii), $\tau_{\text{ir}} = 0.5$ ms, $\tau_{\text{id}} = 1$ ms. In simulations of the spatial model with broad inhibitory projection (Fig. 4Di-Diii), $\alpha_{\text{e}}^{(3)} = 0.1$, $\alpha_{\text{i}}^{(3)} = 0.2$. Other parameters are not changed.

All simulations were performed on the CNBC Cluster in the University of Pittsburgh. All simulations were written in a combination of C and Matlab (Matlab R 2015a, Mathworks). The differential equations of the neuron model were solved using forward Euler method with time step 0.01 ms.

**Experimental methods** Each of the two datasets (recordings from V4 and recordings from V1 and MT) collected from two different rhesus monkeys as they performed an orientation-change detection task. All animal procedures were in accordance with the Institutional Animal Care and Use Committee of Harvard Medical School, University of Pittsburgh and Carnegie Mellon University.

For analysis in Fig. 1Bi and Fig. 4Ai-Aiii, data was collected with two microelectrode arrays implanted bilaterally in area V4 (*1*). We use trials with correct detection for analysis and the first and last stimulus presentations in each trial were not analyzed, to prevent transients

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

due to stimulus appearance or change from affecting the results. Spike counts during the sustained response (120 - 260 ms after stimulus onset) are considered for the correlation and factor analysis. Neurons recorded from either the left or right hemisphere in one session are treated separately. Two sessions from the original study were excluded due to inadequate trials for factor analysis. There are a total of 42,103 trials for 72,765 pairs from 72 recording sessions.

For analysis in Fig. 1Ci, data was collected with one microelectrode array implanted in area V1 and a single electrode or a 24-channel linear probe inserted into MT (*2*). The analysis was performed on responses to 100% contrast stimulus presentations during correct trials before direction change with the exception of the first stimulus presentation. Spike counts are measured 30230 ms after stimulus onset for V1 and 50250 ms after stimulus onset for MT to account for the average visual latencies of neurons in both areas. There are a total of 34,404 V1 pairs from 42 sessions in each attention condition, 1631 V1-MT pairs from 32 recording sessions and 270 MT-MT pairs from 31 sessions.

**Statistical methods** To compute the noise correlation of each simulation, 500 neurons were randomly sampled from the excitatory population of Layer 3 and Layer 2 within a [0, 0.5]x[0, 0.5] square (considering periodic boundary condition). Spike counts were computed using a sliding window of 200 ms with 1 ms step size and the Pearson correlation coefficients were computed between all pairs. Neurons of firing rates less than 2 Hz were excluded from the computation of correlations. In Fig. 3C,D, for each $\mu_i$ there were 50 simulations and each simulation was 20 sec long. Connectivity matrices and the initial states of each neuron's membrane potential were randomized in each simulation. The first 1 second of each simulation was excluded from the correlation analysis. Standard error was computed based on the mean correlations of each simulation. For simulations of Fig. 2D, there was one simulation of 20 seconds per $\tau_{\mathrm{id}}$ and the connectivity matrices were randomized for each simulation. To compute the noise correlation, 1000 neurons were randomly sampled in the excitatory population of Layer 2 within a [0, 0.5]x[0, 0.5] square. Correlations are computed between firing rates that are smoothed with a Gaussian window of width 10 ms.

To study the chaotic population firing rate dynamics of Layer 3 (Fig. 5), we fixed the spike trains realizations from Layer 1 neurons, the membrane potential states of the Layer 2 neurons and all connectivity matrices. Only the initial membrane potentials of Layer 3 neurons were randomized across trials. There were 10 realizations of Layer 1 and Layer 2, each of which was 20 sec long. For each simulation of Layer 2, 20 repetitions with different initial conditions were simulated for Layer 3. The connectivity matrices in Layer 3 were the same across the 20 repetitions but different for each realization of Layer 1 and Layer 2. The realizations of Layer 1 and Layer 2 and the connectivity matrices were the same for the attended and unattended states. Trial-to-trial variance of Layer 3 population rates (Fig. 5E) was the variance of the mean population rates of the Layer 3 excitatory population, smoothed by a 200 ms rectangular filter, across the 20 repetitions. The first second of each simulation was discarded.

Factor analysis assumes spike counts of $n$ simultaneously recorded neurons $x \in \mathbb{R}^{n\times 1}$ is a

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

multi-variable Gaussian process

$$x \sim \mathcal{N}(\mu, LL^T + \Psi)$$

where $\mu \in \mathbb{R}^{n\times 1}$ is the mean spike counts, $L \in \mathbb{R}^{n\times m}$ is the loading matrix of the $m$ latent variables and $\Psi \in \mathbb{R}^{n\times 1}$ is a diagonal matrix of independent variances for each neuron. We choose $m = 5$ and compute the eigenvalues of $LL^T$, $\lambda_i$ $(i = 1, 2, \ldots, 5)$, ranked in descending order. We compute the residual covariance after subtracting the first mode as

$$Q = \text{Cov}(x, x) - L_1 \times L_1'$$

where $\text{Cov}(x, x)$ is the raw covariance matrix of $x$ and $L_1$ is the loading matrix when fitting with $m = 1$. The mean raw covariance and residual (Fig. 4Aiii-Diii) are the mean of the off-diagonal elements of $\text{Cov}(x, x)$ and $Q$, respectively. When applying factor analysis on model simulations (Fig. 4B-D), we randomly selected 50 excitatory neurons from Layer 3, whose firing rates were larger than 2 Hz in both the unattended and attended states. There were 10 non-overlapping sampling of neurons and we applied factor analysis on each sampling of neuron spike counts. There were 15 simulations with fixed connectivity matrices, each of which was 20 seconds long. Spike trains were truncated into 140-ms spike count window with a total of 2025 counts per neuron. In simulations with fast inhibition (Fig. 4Ci-Ciii) and broad inhibitory projection (Fig. 4Di-Diii), the connectivity matrices were the same as those in simulations of the original model (Fig. 4Bi-Biii).

# Supplementary Text

**Hidden variable model.** First, we consider the attentional effect on noise correlations in one cortical area. Let $R$ be the response variable of the neurons in that area and $H$ be the external source of variability which projects to $R$ with strength $\beta$; $R = X + \beta H$. Here we take the variance of $H$, $\text{Var}^{\alpha}(H)$, to be dependent on the attentional state $\alpha \in \{U, A\}$, and for now $X$ is an attention independent source of fluctuating input. Denote $P_H = \frac{\beta^2 \text{Var}^U(H)}{\text{Var}^U(R)} = \frac{\beta^2 \text{Var}^U(H)}{\text{Var}(X)+\beta^2 \text{Var}^U(H)}$ as the influence of $H$ on $R$ $(0 < P_H < 1)$. Then the constraint on $H$ is:

$$\frac{\underset{U-A}{\Delta}\text{Var}(H)}{\text{Var}^U(H)} = \frac{1}{P_H}\frac{\underset{U-A}{\Delta}\text{Var}(R)}{\text{Var}^U(R)}. \tag{6}$$

Here $\underset{U-A}{\Delta}\text{Var}(H) = \text{Var}^U(H) - \text{Var}^A(H)$ (same for $\underset{U-A}{\Delta}\text{Var}(R)$). The population data provide values for $\underset{U-A}{\Delta}\text{Var}(R)/\text{Var}^U(R)$. In Fig. 1Biii (main text) the change in $\text{Var}(H)$, $\underset{U-A}{\Delta}\text{Var}(H)/\text{Var}^U(H)$, is plotted as a function of the influence of $H$ on $R$, $P_H$, with the V4 data (Fig. 1Bi) determining $\underset{U-A}{\Delta}\text{Var}(R)/\text{Var}^U(R) = 0.3$.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we consider the correlation between two cortical areas. Let $R$ be the neural response from MT and $X$ be the neural response from V1. Suppose all the variability in $R$ is from $X$ and the transfer function of $X$ to $R$ can be linearly approximated as $\delta R = \alpha \delta X$, then

$$\begin{aligned} \text{Var}(R) &= \alpha^2 \text{Var}(X), \\ \text{cov}(R, X) &= \alpha \text{Var}(X). \end{aligned}$$

which gives

$$\text{Var}(R) = \text{cov}(R, X)^2 / \text{Var}(X). \tag{7}$$

Hence any decrease in $\text{Var}(R)$ by attention predicts a decrease in $\text{cov}(R, X)$, which is in contradiction with the electrophysiological recordings (*2*) (Fig. 1Bi and Ci).

Assume a hidden source of variability, $H$, that projects to both $R$ and $X$ with strengths $\beta$ and $\kappa$, respectively. Specifically, $R = \alpha X + \beta H$ and $X = X_0 + \kappa H$, where $\text{cov}(X_0, H) = 0$. Suppose $\text{Var}(X) = 1$ and $\beta$ and $\kappa$ are attention independent, then

$$\begin{aligned} \text{Var}(R) &= \alpha^2 + \left(\beta^2 + 2\beta\kappa\alpha\right) \text{Var}(H), \\ \text{cov}(R, X) &= \alpha + \beta\kappa \text{Var}(H). \end{aligned}$$

In order to have an attention-mediated simultaneous reduction in $\text{Var}(R)$ and an increase in $\text{cov}(R, X)$ we need $\alpha$ increases and $\text{Var}(H)$ decreases with attention. The relative change in $\text{cov}(R, X)$ by attention is

$$\frac{\underset{A-U}{\Delta} \text{cov}(R, X)}{\text{cov}^U(E, X)} = \frac{\underset{A-U}{\Delta} \alpha - \beta\kappa \underset{U-A}{\Delta} \text{Var}(H)}{\alpha_U + \beta\kappa \text{Var}^U(H)}. \tag{8}$$

An attention-mediated increase of the correlation between V1 ans MT implies $\underset{A-U}{\Delta} \text{cov}(R, X) > 0$, which gives

$$\underset{A-U}{\Delta} \alpha > \beta\kappa \underset{U-A}{\Delta} \text{Var}(H). \tag{9}$$

The reduction in $\text{Var}(R)$ by attention is

$$\begin{aligned} \underset{U-A}{\Delta} \text{Var}(R) &= -\left(\alpha_A^2 - \alpha_U^2\right) + \beta^2 \underset{U-A}{\Delta} \text{Var}(H) + 2\beta\kappa \left(\alpha_U \text{Var}^U(H) - \alpha_A \text{Var}^A(H)\right) \\ &= -\left(2\alpha_U + \underset{A-U}{\Delta} \alpha\right) \underset{A-U}{\Delta} \alpha + \beta^2 \underset{U-A}{\Delta} \text{Var}(H) + 2\beta\kappa\alpha_U \underset{U-A}{\Delta} \text{Var}(H) - 2\beta\kappa \text{Var}^A(H) \underset{A-U}{\Delta} \alpha \\ &= \beta^2 \underset{U-A}{\Delta} \text{Var}(H) - 2\alpha_U \underset{A-U}{\Delta} \text{cov}(R, X) - \left(\underset{A-U}{\Delta} \alpha\right)^2 - 2\beta\kappa \text{Var}^A(H) \underset{A-U}{\Delta} \alpha \end{aligned}$$

Hence the relative reduction in $\text{Var}(R)$ is

$$\frac{\underset{U-A}{\Delta} \text{Var}(R)}{\text{Var}^U(R)} = \frac{\underset{U-A}{\Delta} \text{Var}(H)}{\text{Var}^U(H)} P_H - 2\alpha_U \frac{\underset{A-U}{\Delta} \text{cov}(R, X)}{\text{cov}^U(R, X)} \frac{\text{cov}^U(R, X)}{\text{Var}^U(R)} - \frac{\left(\underset{A-U}{\Delta} \alpha\right)^2 + 2\beta\kappa \text{Var}^A(H) \underset{A-U}{\Delta} \alpha}{\text{Var}^U(R)} \tag{10}$$

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The second term from the RHS of Eq. (10) is

$$2\alpha_U \frac{\text{cov}^U(R,X)}{\text{Var}^U(R)} \frac{\underset{A-U}{\Delta}\text{cov}(E,X)}{\text{cov}^U(E,X)} = \frac{2\alpha_U^2 + 2\beta\kappa\alpha_U \text{Var}^U(H)}{\alpha_U^2 + (\beta^2 + 2\beta\kappa\alpha_U)\,\text{Var}^U(H)} \frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)}$$
$$> (1-P_H)\frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)}$$

With inequality (9), the third term from the RHS of Eq. (10) is

$$\frac{\left(\underset{A-U}{\Delta}\alpha\right)^2 + 2\beta\kappa\text{Var}^A(H)\underset{A-U}{\Delta}\alpha}{\text{Var}^U(R)} = \frac{\underset{A-U}{\Delta}\alpha\left(\underset{A-U}{\Delta}\alpha + 2\beta\kappa\text{Var}^A(H)\right)}{\beta^2\text{Var}^U(H)}P_H$$
$$> \frac{\left(\beta\kappa\underset{U-A}{\Delta}\text{Var}(H)\right)\left(\beta\kappa\underset{U-A}{\Delta}\text{Var}(H) + 2\beta\kappa\text{Var}^A(H)\right)}{\beta^2\text{Var}^U(H)}P_H$$
$$= \frac{\beta^2\kappa^2\underset{U-A}{\Delta}\text{Var}(H)\left(\text{Var}^U(H) + \text{Var}^A(H)\right)}{\beta^2\text{Var}^U(H)}P_H$$
$$> \kappa^2\text{Var}^U(H)\frac{\underset{U-A}{\Delta}\text{Var}(H)}{\text{Var}^U(H)}P_H$$

Hence,

$$\frac{\underset{U-A}{\Delta}\text{Var}(R)}{\text{Var}^U(R)} < \left(1-\kappa^2\text{Var}^U(H)\right)\frac{\underset{U-A}{\Delta}\text{Var}(H)}{\text{Var}^U(H)}P_H - \frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)}(1-P_H)$$

which gives the constraint on $H$ as

$$\frac{\underset{U-A}{\Delta}\text{Var}(H)}{\text{Var}^U(H)}P_H > \frac{\underset{U-A}{\Delta}\text{Var}(R)}{\text{Var}^U(R)}\frac{1}{1-\kappa^2\text{Var}^U(H)} + \frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)}\frac{1-P_H}{1-\kappa^2\text{Var}^U(H)}. \tag{11}$$

where $0 < 1-\kappa^2\text{Var}^U(H) = \frac{\text{Var}^U(X_0)}{\text{Var}(X)} < 1$, since $\text{Var}(X) = \text{Var}(X_0) + \kappa^2\text{Var}(H) = 1$. Therefore, the lower bound on $\frac{\underset{U-A}{\Delta}\text{Var}(H)}{\text{Var}^U(H)}P_H$ increases with the relative increase in $\text{cov}(R,X)$, $\frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)}$, and the projection strength of $H$ on $X$, $\kappa$.

Fig. 1Biii is plotted using Eq. (6) and Fig. 1Ciii is plotted using Eq. 11 with $\frac{\underset{A-U}{\Delta}\text{cov}(R,X)}{\text{cov}^U(R,X)} = 1$ and $\kappa^2\text{Var}(H)$ ranges from 0 to 0.5 for different curves.

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Neural field model and stability analysis** We use a two dimensional neural field model to describe the dynamics of population rate. Consider the neural field equations

$$\tau_\alpha \frac{\partial r_\alpha(x,t)}{\partial t} = -r_\alpha + \phi(w_{\alpha e} * r_e + w_{\alpha i} * r_i + \mu_\alpha) \tag{12}$$

where $r_\alpha(x,t)$ is the firing rate of neurons in population $\alpha = e, i$ near spatial coordinates $x \in [0,1] \times [0,1]$. The symbol $*$ denotes convolution in space, $\mu_\alpha$ is a constant external input and $w_{\alpha\beta}(x) = \overline{w}_{\alpha\beta} g(x;\sigma_\beta)$ where $g(x;\sigma_\beta)$ is a two-dimensional wrapped Gaussian with width parameter $\sigma_\beta$, $\beta = e, i$ (*3*). The transfer function is a threshold-quadratic function, $\phi(x) = [x^2]_+$. The timescale of synaptic and firing rate responses are implicitly combined into $\tau_\alpha$. In networks with approximate excitatory-inhibitory balance, rates closely track synaptic currents (*4*), so $\tau_\alpha$ represents the synaptic time constant of population $\alpha = e, i$.

For constant inputs, $\mu_e$ and $\mu_i$, there exists a spatially uniform fixed point, which was computed numerically using an iterative scheme (*3*). Linearizing around this fixed point in Fourier domain gives a Jacobian matrix at each spatial Fourier mode (*3, 5*)

$$J(\vec{n}) = \begin{bmatrix} \left(-1 + g_e \widetilde{w}_{ee}(\vec{n})\right)/\tau_e & g_e \widetilde{w}_{ei}(\vec{n})/\tau_e \\ g_i \widetilde{w}_{ie}(\vec{n})/\tau_i & \left(-1 + g_i \widetilde{w}_{ii}(\vec{n})\right)/\tau_i \end{bmatrix}.$$

where $\vec{n} = (n_1, n_1)$ is the two-dimensional Fourier mode, $\widetilde{w}_{\alpha\beta}(\vec{n}) = \overline{w}_{\alpha\beta} \exp(-2\|\vec{n}\|^2 \pi^2 \sigma_\beta^2)$ is the Fourier coefficient of $w_{\alpha\beta}(x)$ with $\|\vec{n}\|^2 = n_1^2 + n_2^2$ and $g_a$ is the gain, which is equal to $\phi'(r_\alpha)$ evaluated at the fixed point. The fixed point is stable at Fourier mode $\vec{n}$ if both eigenvalues of $J(\vec{n})$ have negative real part. A Hopf bifurcation occurs when complex eigenvalues with positive real part emerge at the uniform Fourier mode, $\vec{n} = (0,0)$. A Turing-Hopf bifurcation occurs when complex eigenvalues with positive real part emerge at any non-uniform Fourier mode, $\vec{n} \neq (0,0)$. Note that stability only depends on the wave number, $k = \|\vec{n}\|$, so Turing-Hopf instabilities always occur simultaneously at all Fourier modes with the same wave number.

Simulations were performed by discretizing space uniformly into a $100 \times 100$ grid and using a forward Euler method to solve Eq. (12). Parameters were $\overline{w}_{ee} = 80$, $\overline{w}_{ei} = -160$, $\overline{w}_{ie} = 120$, and $\overline{w}_{ii} = -200$, $\mu_e = 0.48$ and $\mu_i = 0.32$. For the stability analysis in Fig. 4E, $\tau_i$ varies from 2.5 ms to 25 ms, $\sigma_i$ varies from 0.05 to 0.2, and $\tau_e = 5$ ms and $\sigma_e = 0.1$. Destabilizing the inhibitory population ($\mu_I = 0.5$) expands the stable region (Fig. 4E, black dashed).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Supplementary figures

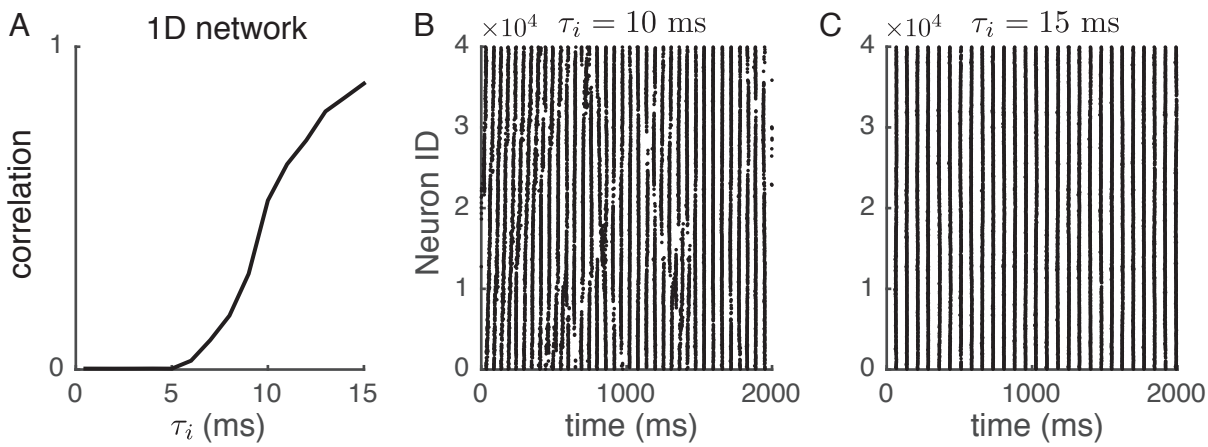


Figure S1: Supplementary Figure S1: (A) One-dimensional spatial model shows rapid increase in mean pairwise correlation with increasing time scale of inhibition (compare with Fig. 2D). (B) Example rasters of network activity when $\tau_i = 10$ ms. (C) Same as B with $\tau_i = 15$ ms. Parameters of the one dimensional model are the same as those in the two-dimensional spatial model in Fig. 2, except that neurons are ordered on interval [0, 1].

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

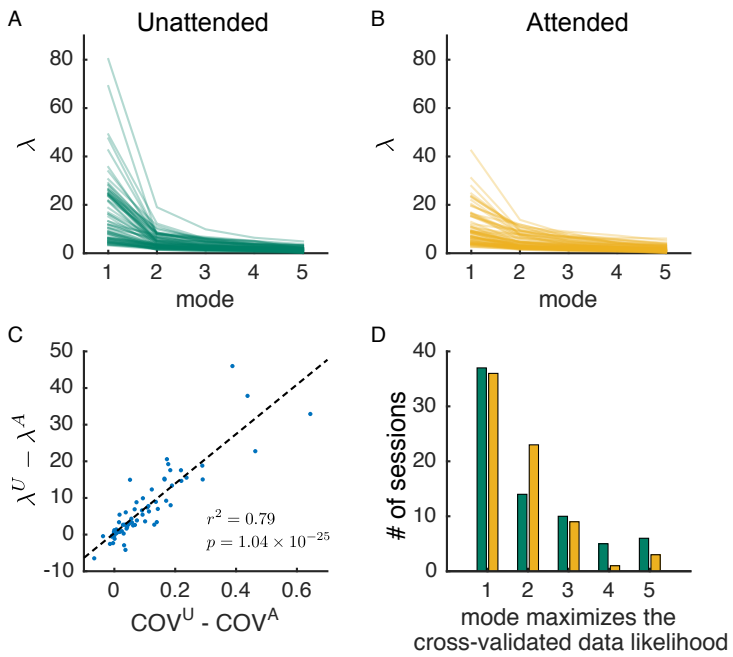


Figure S2: Supplementary Figure S2: Factor analysis of the multi-electrode recordings from V4 (*1*). (A) Eigenvalues of the first five modes of the shared component of the spike count covariance matrix. Each line is for data from each recording session (72 in total). (B) Same as A for the attended state. (C) The difference of the largest eigenvalue and the difference of mean covariance between unattended and attended states are correlated. (D) Histogram of the modes that maximize the cross-validated data likelihood across sessions. More details see Experimental methods and Statistical methods.

# Supplementary Movie Captions

Movie S1: Spiking activities of a spatially ordered network with fast inhibition (Fig. 2Aii). Each black dot indicates that the neuron at spatial position (x; y) fired within one millisecond of the time stamp shown on top.

Movie S2: Same as Movie S1 for a spatially ordered network with slow inhibition (Fig. 2Aiv)

# References

1. M. Cohen, J. Maunsell, *Nature neuroscience* **12**, 1594 (2009).

2. D. A. Ruff, M. R. Cohen, *Journal of Neuroscience* **36**, 7523 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/217976; this version posted November 11, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. R. Rosenbaum, B. Doiron, *Physical Review X* **4**, 021039 (2014).

4. A. Renart, *et al.*, *Science* **327**, 587 (2010).

5. R. Pyle, R. Rosenbaum, *Physical Review Letters* **118**, 018103 (2017).