The mechanics of correlated variability in segregated cortical excitatory subnetworks

Alex Negrón[1,3,*], Matthew P. Getz[2,3,*,‡], Gregory Handy[2,3,†], Brent Doiron[2,3,†,‡]

1 Department of Applied Mathematics, Illinois Institute of Technology
2 Departments of Neurobiology and Statistics, University of Chicago
3 Grossman Center for Quantitative Biology and Human Behavior, University of Chicago

***These authors contributed equally.**
†These authors contributed equally.
‡mgetz@uchicago.edu, bdoiron@uchicago.edu

Abstract

Understanding the genesis of shared trial-to-trial variability in neural activity within sensory cortex is critical to uncovering the biological basis of information processing in the brain. Shared variability is often a reflection of the structure of cortical connectivity since this variability likely arises, in part, from local circuit inputs. A series of experiments from segregated networks of (excitatory) pyramidal neurons in mouse primary visual cortex challenge this view. Specifically, the across-network correlations were found to be larger than predicted given the known weak cross-network connectivity. We aim to uncover the circuit mechanisms responsible for these enhanced correlations through biologically motivated cortical circuit models. Our central finding is that coupling each excitatory subpopulation with a specific inhibitory subpopulation provides the most robust network-intrinsic solution in shaping these enhanced correlations. This result argues for the existence of excitatory-inhibitory functional assemblies in early sensory areas which mirror not just response properties but also connectivity between pyramidal cells.

1 Introduction

Determining a structure – function relationship in a cortical circuit is a central goal in many neuroscience research programs. While the trial averaged responses of a network to a fixed stimulus or repeated behavior does give some information about the underlying circuit, the dynamic or trial-to-trial fluctuations of neuronal activity provides another important glimpse into network structure (Urai et al., 2022). Such neuronal variability is a salient feature of cortical responses (Faisal et al., 2008), and of particular interest is how that variability is distributed over a population of neurons (Cohen and Kohn, 2011). The shared fluctuations of a pair of neurons, termed *noise correlations*, are often thought to reflect the circuit structure of the network within which the neuron pair is embedded (Doiron et al., 2016; Ocker et al., 2017). Understanding how neural variability is shaped by the connections and local circuit dynamics can provide rich insight into the structure and function of cortical circuits.

An early hope was that pairwise correlations in neuronal activity could be used to infer the underlying connectivity in a straightforward fashion (Mishchencko et al., 2011;

Roudi et al., 2015). Indeed, experiments in the mouse primary visual cortex (V1) demonstrated that the magnitude of the pairwise correlation between two pyramidal cells increases with their probability of connection (Ko et al., 2011; Cossell et al., 2015). Theoretically, this result can be thoroughly explained in a weakly coupled excitatory network, where correlations are dominantly determined by direct, monosynaptic connections. However, recent experiments investigating the functional properties of two distinct subpopulations of pyramidal cells in mouse V1 complicate this narrative (Kim et al., 2018b). These subpopulations project to separate downstream higher visual areas and are inter-connected with lower probability than that of randomly sampled pyramidal cells within V1. Despite this weak connectivity, it was found that the correlations between these distinct subpopulations were much higher than predicted by their sparse inter-connectivity. In fact, the magnitude of the correlated variability across the two subpopulations approached that between any randomly chosen pair of excitatory neurons. In this same vein, another experiment examining callosal projection neurons in mouse V1 found that these cells also cluster and connect more strongly as a class (Hagihara et al., 2021). Yet their correlated variability is similar when comparing within-class and out-of-class, again illustrating that significant, positive correlations can persist in the absence of direct strong connections. In total, these results are at odds with previous intuition, namely that this anatomical segregation would correspond to a functional one as well.

Theoretical work has also highlighted how the simplistic structure-dynamics relationship originally put forth can break down. For example, it has been shown that inferring connectivity from activity becomes difficult as recurrent connection strengths grow and inhibition is required to stabilize the network (Das and Fiete, 2020; Biswas and Fitzgerald, 2022). Most notably, a densely connected network with strong synaptic weights that exists in the so-called balanced state, a robust parameter regime where excitatory and inhibitory inputs to a neuron largely cancel out, results in near-zero average correlations (Renart et al., 2010; Rosenbaum et al., 2017). Nonetheless, progress has been made in overcoming these difficulties, with studies having developed methods for linking connectivity motifs to the structure of correlations in arbitrarily large networks (Pernice et al., 2011; Trousdale et al., 2012; Ocker et al., 2017)).

In this work, we seek to apply some of these techniques to characterize the neural circuit properties which could explain the significant positive shared variability across segregated cortical subpopulations observed in Kim et al. (2018b). With the use of mean field circuit models we show that the solution depends on the dynamical regime of the circuit, and relies on the structure of inhibition. In a weakly coupled regime, correlations can be characterized through inheritance from outside sources, or increased through shared inhibitory inputs. By contrast, in a strongly coupled regime, shared inhibition would largely act to anticorrelate activity across the populations. Critically, we show that this anticorrelation can be mitigated if inhibition is similarly clustered with excitation, forming instead excitatory-inhibitory assemblies. Additionally, this regime of strongly coupled dynamics with clustered inhibition provides the most robust solution space to explain the elevated correlations. This prediction further suggests that other apparent correlation conundrums could be solved by supplementing excitatory recordings with activity from inhibitory neurons.

2 Results

2.1 Segregated synaptic wiring does not produce segregated functional responses

Our work is motivated by an apparent inconsistency in a series of experimental studies exploring the relation between the recurrent circuitry and functional responses of neuronal populations in sensory neocortex. Ko et al. (2011) and Cossell et al. (2015) used a combination of *in vivo* population imaging and *in vitro* electrophysiology to show that the activity correlations between pairs of pyramidal neurons in mouse primary visual cortex (V1) increase monotonically with the probability of there existing synaptic connections between them. Later work from the same group (Kim et al., 2018b) investigated two excitatory populations in mouse V1: neurons that are either anterolateral (AL)- or posteromedial (PM)- projecting. Despite neurons being in close spatial proximity to each other, these neuronal subpopulations exhibit high within group connectivity (prob. AL $\leftrightarrow$ AL connection ~ 0.21, prob. PM $\leftrightarrow$ PM connection ~ 0.18) and low between group connectivity (prob. AL $\rightarrow$ PM connection ~ 0.04, prob. PM $\rightarrow$ AL connection ~ 0.05). To streamline our presentation we will label these two populations E_1 and E_2 (Fig. 1A). Given the low connection probability between E_1 and E_2 and the established relation between connectivity and activity correlations shown in Ko et al. (2011) and Cossell et al. (2015), one would predict that the degree of correlations between the activities of E_1 and E_2 would be low (Fig. 1B, held out light green square; from Kim et al. (2018b) we estimate this value to be approximately -0.05). However, Kim et al. (2018b) reported substantially higher than predicted mean E_1 and E_2 correlations (Fig. 1B, dark green square; Kim et al. (2018b) measured it to be about 0.027, close to the within-group values of about 0.035-0.04). In total, while pyramidal neurons in mouse V1 projecting to distinct targets show segregated synaptic connectivity, the degree of functional segregation between these subpopulations is below what is expected.

The central goal of our study is to put forth a circuit-based model framework that can robustly and self-consistently account for both of these experimental observations. It is important to note that Kim et al. (2018b) only considered total correlations (of the raw neural activity traces) in computing this expected correlation value. However, given the similarities observed in the signal and noise correlation structure in both this and previous studies (Ko et al., 2011; Kim et al., 2018b; Hagihara et al., 2021), we focus here on noise correlations which relate more directly to the underlying structure of connectivity (Ocker et al., 2017).

2.2 A circuit model of fluctuations in segregated subpopulations

To study the structure of correlations in anatomically segregated networks and investigate the possible mechanisms responsible for the unexpectedly enhanced correlations, we consider a phenomenological dynamic mean field model for the aggregate activity of each neural population (Renart et al., 2004; Getz et al., 2022; Kanashiro et al., 2017). Assuming that the network has a steady state solution ($\mathbf{r}_{\text{ss}}$), the linearized dynamics of population A around this equilibrium are given by (see Section 5 for additional details):

$$\tau_A \frac{d\Delta r_A}{dt} = -\Delta r_A + \sum_B W_{AB}\Delta r_B + \sigma_A \left[\sqrt{1-c}\cdot\xi_A(t) + \sqrt{c}\cdot\xi_S(t)\right] \quad (1)$$

where $\Delta r_A = r_A - r_{\text{ss},A}$, τ_A is a time constant, and W_{AB} is the effective strength of connections from population B to A. For the purely excitatory network A and B range

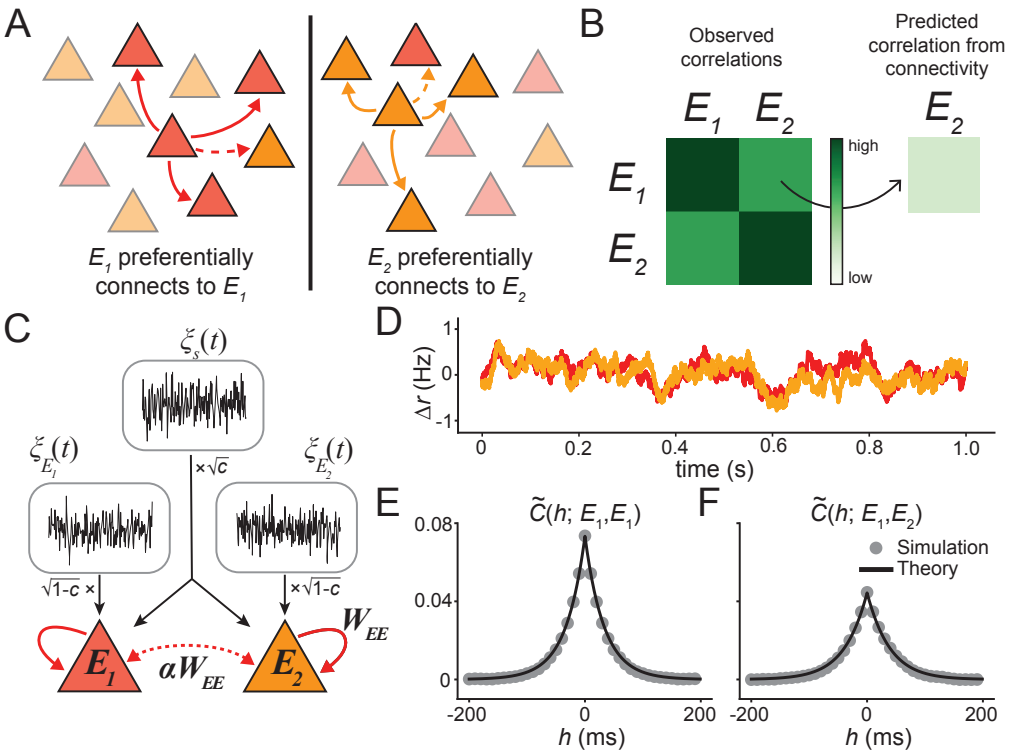


Figure 1. Mean field model of segregated E populations. A: Illustration of experimentally observed connectivity motif; the red (E_1) and orange (E_2) populations connect with lower probability than average. **B:** Schematic of main experimental observations: $E_1 - E_2$ correlations were higher than would be predicted from their low connectivity. **C:** Model schematic. Black traces and arrows denote noise sources. Red arrows indicate excitatory recurrent connections where the dashed line connotes weakened connection strength. Feedforward stimulus drive omitted for clarity. **D:** Example realization of network activity to a sustained, fixed stimulus. Colors as in (A). **E:** E_1 auto-correlation function and **F:** $E_1 - E_2$ cross-correlation function for the illustrated rate traces. For panels D, E and F: $c = 0.5$.

over E_1 and E_2; when inhibitory connections are included in later sections A and B will include those as well. The stochastic processes $\xi_A(t)$ and $\xi_S(t)$ represent private and shared global fluctuations, respectively, modelling stochastic inputs that are external to the network. $\xi_A(t)$ and $\xi_S(t)$ are taken to be independent Gaussian processes with $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \delta(t - t')$. The parameter $c \in [0, 1]$ scales the proportion of shared noise relative to private noise (Fig. 1C), while $\sigma_A > 0$ represents the total intensity of the external fluctuations given to population A.

We make two assumptions: 1) the network has a stable solution r_{ss} about which the population dynamics fluctuate (Fig. 1D), and 2) connections within and inputs to the network are symmetric across the two E populations, with $W_{E_1E_1} = W_{E_2E_2} = W_{EE}$, $W_{E_1E_2} = W_{E_2E_1} = \alpha W_{EE}$, and $\sigma_{E_1} = \sigma_{E_2} = \sigma$. Note that parameter $0 < \alpha \ll 1$ represents the degree to which the inter-population connections are weaker than the within-population connections (Fig. 1C). Since the system of recurrently coupled stochastic differential equations in Eq. 1 is a multi-dimensional Ornstein-Uhlenbeck (OU) process, we can derive (see Section 5) an analytical formula for its stationary autocovariance function

$$\widetilde{\mathbf{C}}(h) = \langle \Delta\mathbf{r}(t), \Delta\mathbf{r}(t+h) \rangle,$$

which agrees well with numerical simulations (Fig. 1E, F). Further, and of particular interest in this work, is the *long-time covariance matrix* defined as

$$\mathbf{C} := \int_{-\infty}^{\infty} \widetilde{\mathbf{C}}(h) dh.$$

This may be expressed (see Methods 5.2) as

$$\mathbf{C} = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{D}\left[(\mathbf{I} - \mathbf{W})^{-1}\mathbf{D}\right]^{\top}, \tag{2}$$

where $\mathbf{W}$ is a matrix of effective connection strengths and $\mathbf{D}$ is a matrix that scales the fluctuations. We define the correlations between E_1 and E_2 as

$$\mathrm{Corr}(E_1, E_2) := \frac{C_{E_1E_2}}{\sqrt{C_{E_1E_1}C_{E_2E_2}}} = \frac{C_{E_1E_2}}{C_{E_1E_1}}, \tag{3}$$

where C_{AB} is an element of $\mathbf{C}$ and the second equality follows by the assumed symmetry in the system. This framework enables us to formalize the motivating question of our study: what are the mechanisms that enable higher than expected correlations across anatomically segregated populations? For the sake of specificity, we choose the threshold $\mathrm{Corr}(E_1, E_2) > 0.6$ as an approximation of the ratio of mean across-population to within-population noise correlations in Kim et al. (2018b).

2.3 Inheritance model of correlations between weakly coupled excitatory populations

We begin by exploring how the strength of recurrent excitation (W_{EE}) and the proportion of fluctuations that are shared (c) shape correlations between the segregated E populations. In this section, to ensure that the network admits a stable activity solution we require $W_{EE} < 1$, else recurrent excitation would lead to runaway activity. Note that while we allow W_{EE} to vary, we maintain the concept of segregated populations by keeping α small and fixed. We find that while increasing W_{EE} leads to moderate increases in $\mathrm{Corr}(E_1, E_2)$, a much more significant increase occurs by increasing c (Fig. 2A).

To better understand the underlying mechanisms responsible for these higher correlations within this parameter regime (i.e., to the right of the pink line in Fig. 2A), we perform a pathway expansion of the covariance matrix Eq. 2. Since the steady state emitted by the system in Eq. 1 is stable, the term $(\mathbf{I} - \mathbf{W})^{-1}$ can be expanded as a series. This allows us to write Eq. 2 as (see Methods 5.3)

$$\mathbf{C} = \sum_{n=0}^{\infty}\left[\sum_{i=0}^{n}\mathbf{W}^{n-i}\mathbf{D}\mathbf{D}^{\top}\left(\mathbf{W}^{\top}\right)^{i}\right], \tag{4}$$

where each term in the inner sum corresponds to an n^{th}-order path through the network. Writing out the first three terms of this sum for the cross-covariance yields

$$C_{E_1E_2} = \sigma^2\left[c + (2c + 2\alpha)W_{EE} + (3(1+\alpha^2)c + 6\alpha)W_{EE}^2\right] + \mathcal{O}(W_{EE}^3).$$

Rewriting this equation as

$$C_{E_1E_2} = \sigma^2[\underbrace{c \cdot (1 + 2W_{EE} + 3(1+\alpha^2)W_{EE}^2)}_{(1)} + \underbrace{(2\alpha W_{EE} + 6\alpha W_{EE}^2)}_{(2)}] + \mathcal{O}(W_{EE}^3) \tag{5}$$

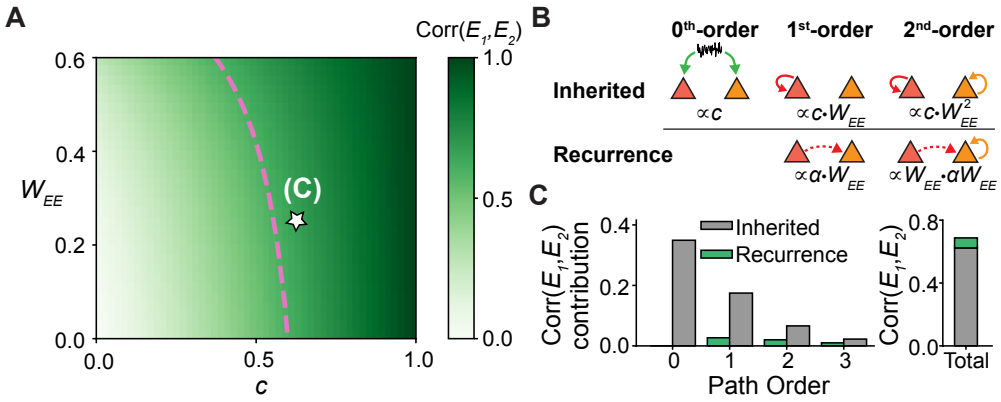


Figure 2. Highly correlated regime in weakly coupled excitatory network relies on correlated feedforward inputs. A: $\mathrm{Corr}(E_1, E_2)$ as a function of W_{EE} and the magnitude of shared input noise c. Dashed pink line indicates $\mathrm{Corr}(E_1, E_2) = 0.6$, approximating the value reported in Kim et al. (2018b). **B:** Schematic of example synaptic paths through the network, along with their contribution to the cross-covariance, relating to the path expansion Eq. 4. The inherited row refers to correlated paths stemming from correlations in the feedforward input, while the recurrence row arises from the recurrent connections across the populations. **C:** Contributions of paths of given order to networks (left) and the total correlation (right) for the parameters $W_{EE} = 0.25$ and $c = 0.65$ (star from panel (A)). All panels: $\alpha = 0.1$.

reveals that each term contributing to this cross-covariance can be thought of as arising from one of two sources: 1) inherited from the shared correlated input and dependent on the parameter c (Fig. 2B, top), and 2) purely arising from the recurrent connections (Fig. 2B, bottom). We note that the 'propagation' of the inherited contribution to higher-order paths does not only rely on the $E_1 \leftrightarrow E_2$ connections (proportional to $\alpha^n c$). This is because the correlated activity is fed directly into each subpopulation at the 0^{th} order, from which it can propagate into higher-order paths via self loops contained within each population. We emphasize that eliminating the 0^{th} order term (i.e., setting $c = 0$) eliminates all contributions from the inherited global source.

We now utilize this pathway expansion to compare the contributions from feedforward and recurrent mechanisms to the net cross-covariance for an example point lying in the highly correlated regime (Fig. 2A, star; $W_{EE} = 0.25$ and $c = 0.65$). We first note that this series converges quickly and only a few paths significantly contribute to the total correlation (Fig. 2C). The convergence of this series depends directly on the largest eigenvalue of $\mathbf{W}$ (Methods 5.3), namely

$$\lambda_{\max} = W_{EE} \cdot (1 + \alpha),$$

which is small for our choice of parameters. Our numerical results also illustrate that the contribution from the inherited source largely dominates at each order (Fig. 2C; left), and contributes ∼90% of the total cross-correlation (Fig. 2C; right). These results hold qualitatively across this parameter regime, and lead us to conclude that it corresponds to a model in which large shared input fluctuations explain the heightened correlations between the separate E populations. Taken together, we characterize this solution which exhibits enhanced $E_1 - E_2$ correlations as a *feedforward inheritance model*.

However, under the condition where the shared input fluctuations are small, we still lack a potential mechanism for significant positive correlations. To surmount this shortcoming, we first need to extend our model to also include inhibitory populations.

2.4 Weak recurrent excitation with global inhibition

Parsimoniously, we begin by modelling inhibition as a single global population, consistent with observations that inhibition simply connects densely and non-specifically within cortex (Hofer et al., 2011; Packer and Yuste, 2011) (Fig. 3A). To understand the effect of inhibition in this circuit, we explore how the strength of recurrent inhibitory connections ($W_{EI} < 0$ and $W_{IE} > 0$) shape correlations between the excitatory populations in the case when $c = 0$. Assuming W_{EE} remains weak (i.e., $W_{EE} < 1$), we find a large portion of the parameter regime yields negative cross-correlations (Fig. 3B; purple region). However, there is a region that satisfies our correlation condition, namely the dark green region that corresponds to strong $I \to E$ and weak $E \to I$ connections.

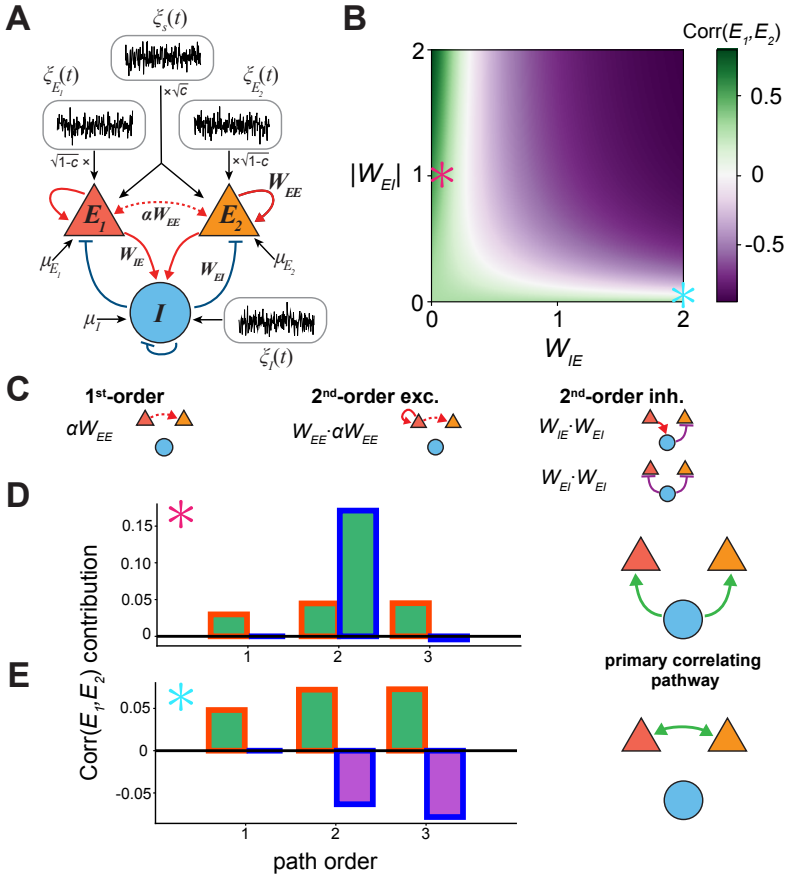


Figure 3. Weakly coupled network. A: Network model schematic as in Figure 1C. Blue lines indicate recurrent inhibitory connections. **B:** $\mathrm{Corr}(E_1, E_2)$ as a function of $|W_{EI}|$ and W_{IE}. **C:** Illustrations of first and second order paths. **D, E:** (Left) Contributions of E (red outlined bars) and I (blue outlined bars) to the net $\mathrm{Corr}(E_1, E_2)$. (Right) Schematic of dominant correlating pathway. Colored stars denote locations in B. Red star: $W_{EI} = -1, W_{IE} = 0.07$; blue star: $W_{EI} = -0.05, W_{IE} = 2$. For all panels $\alpha = 0.15$.

We again make use of a pathway expansion of Eq. 2 to help decipher this observation, this time accounting for the new inhibitory pathways (Fig. 3C). Writing out the expansion to second order in W yields

$$C_{E_1E_2} = \sigma^2[\underbrace{2\alpha W_{EE} + 6\alpha W_{EE}^2}_{\text{exc. paths}} + \underbrace{2W_{EI}W_{IE} + W_{EI}^2}_{\text{inh. paths}}] + \mathcal{O}(W^3), \tag{6}$$

where we have noted the terms involving only the excitatory components and terms which involve paths through the inhibitory population. We first observe that contributions to the cross-covariance due to the excitatory subnetwork at each order are the same as the previous network without the inhibitory connections (Eq. 5 for $c = 0$). This leads us to decompose the total covariance into an excitatory component and an inhibitory component (neglecting the $\mathcal{O}(W^3)$ terms in Eq. 6)

$$C_{E_1E_2} = C^{\text{exc}}_{E_1E_2} + C^{\text{inh}}_{E_1E_2}. \tag{7}$$

As Eq. 6 suggests, depending on the strength of the underlying inhibitory connections, $C^{\text{inh}}_{E_1E_2}$ can either be positive (positively correlating the excitatory subpopulations; Fig. 3B, along W_{EI} axis) or negative (anti-correlating the subpopulations; Fig. 3B, purple region). By contrast, $C^{\text{exc}}_{E_1E_2}$ is clearly bounded below by zero.

Specifically, Eq. 6 reveals a 'tug of war' that can arise early on in the pathway expansion between the $E \to I \to E$ (i.e., $W_{EI}W_{IE} < 0$) and the $I \to E$ (i.e., $W^2_{EI} > 0$) inhibitory pathways. Choosing $|W_{EI}| > W_{IE} \approx 0$, we find that the positive term dominates, and the the inhibitory population acts as a strong correlator of excitatory activity (Fig. 3D). We term this an *inhibitory inheritance model* by analogy to the feedforward inheritance model described above.

On the other hand, when $W_{IE} > |W_{EI}| \approx 0$, the negative term dominates, leading the inhibitory population to weaken the strength of cross-correlations. In this case, the primary correlating source across the excitatory populations are the weak $E_1 \leftrightarrow E_2$ connections (Fig. 3E). But as we noted previously (Fig. 2), this pathway alone is incapable of yielding high cross-correlations without strongly correlated feedforward input.

The regime of weakly coupled neural populations thus permits two solutions for correlating E_1 and E_2 to a sufficiently high degree, both of which can be characterized in terms of inheritance models. Namely, enhanced positive correlations can be inherited from outside sources or from local recurrent inhibition. Nevertheless, the ambiguity in the former solution and the fine-tuning required to achieve the latter solution push us to uncover a more robust mechanism.

2.5 Strong recurrent excitation with global inhibition

Up to this point, by virtue of our assumption that the recurrent excitatory coupling is weak the stability of the equilibrium point was independent of the inhibitory currents. Such a network is commonly referred to as a non-inhibition-stabilized network (non-ISN) (Tsodyks et al., 1997; Ozeki et al., 2009; Sadeh and Clopath, 2021) (see Appendix for additional details). However, recent experimental evidence suggests that mouse cortex operates in the ISN regime, where strong recurrent excitation is tracked and balanced by strong inhibitory feedback (Adesnik, 2017; Sanzeni et al., 2020). Since the ISN regime is known to exhibit sometimes perplexing dynamics, such as the well-studied paradoxical effect (Tsodyks et al., 1997), it is initially unclear how shifting into this parameter regime will shape the correlations under investigation.

In view of this, we now strengthen the recurrent excitatory connections W_{EE} such that our model network lies in the ISN regime. Performing a similar analysis as before (i.e., fixing W_{EE} and W_{II}, while varying W_{EI} and W_{IE}) and assuming that the feedforward inputs are uncorrelated ($c = 0$), we find results that at first glance appear familiar (Fig. 4A). Namely, a portion of the parameter regime results in negative correlations (purple region), with a narrow parameter regime yielding positive correlations (green region). However, unlike the previous network, these correlations are much larger across this band of parameter values, approaching unity as the system loses stability due to the inhibitory feedback becoming too weak to be able to balance out the strong excitation (gray and red-hatched region).

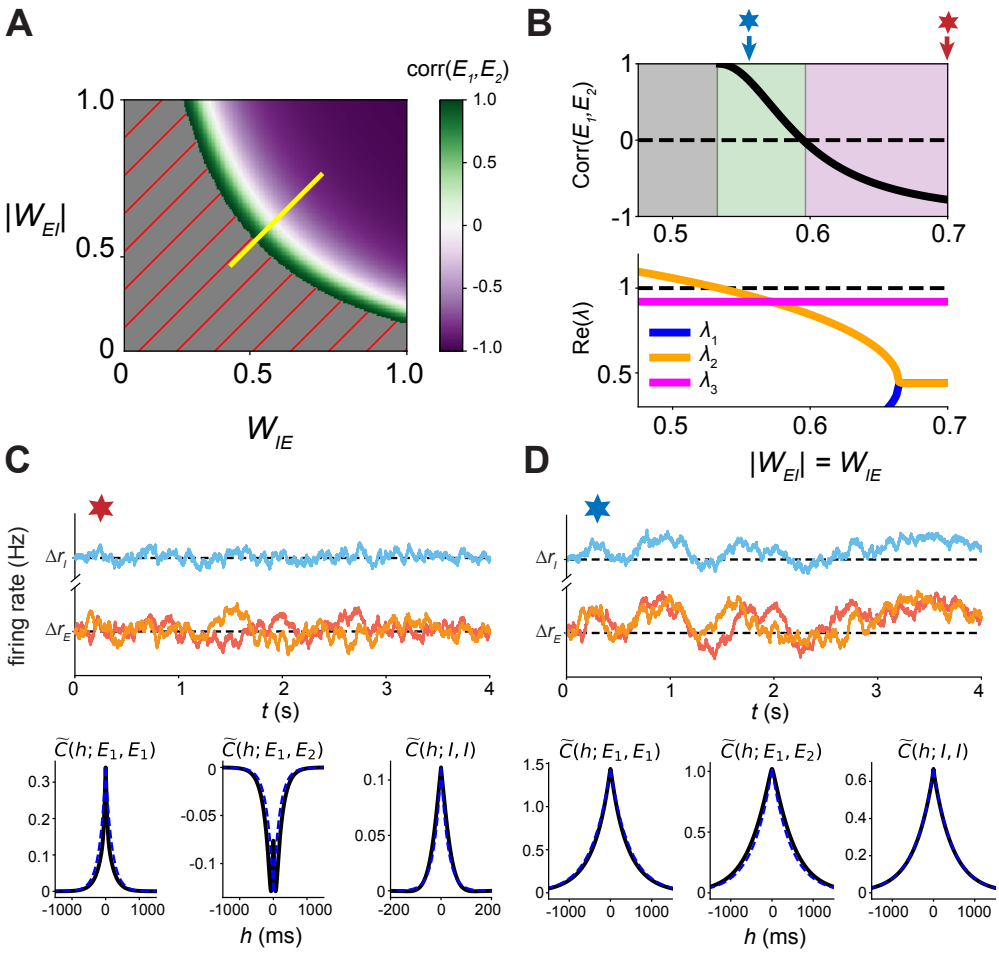


Figure 4. Global inhibition in ISN regime. **A:** Corr(E_1, E_2) as a function of W_{EI}, W_{IE} with $c = 0$. **B:** Top: Corr(E_1, E_2) along the yellow path in A. Gray region: unstable; green region: positive correlations; purple region: negative correlations. Bottom: eigenvalues of the circuit along the yellow path in A. **C, D:** Top: example rate traces (colors as in Fig. 3B). Bottom: auto- and cross-correlation functions computed numerically (black) and theoretically for the dominant timescale (blue dashed). Stars indicate parameter values shown in B. Here, $\alpha = 0.2$.

Unlike the non-ISN regime, where the weak recurrent excitatory connections corresponded with small eigenvalues and quick convergence in our path-expansion, here the eigenvalues of the system lie much closer to the boundary separating stability from instability. As result, many more terms are needed before the series in Eq. 4 converges, complicating its interpretation. Instead, we seek to understand the mechanism driving these high correlations by exploring their apparent connection to the system's stability.

We start by considering the slice of the parameter space where $|W_{EI}| = W_{IE}$ that captures the system's transitions from negative correlations to positive correlation to instability (Fig. 4A, yellow line; Fig. 4B, top). Analysis of the eigenvalues of $\mathbf{W}$ reveals a pair of eigenvalues (λ_1 and λ_2) dependent on the strength of inhibitory connections and another eigenvalue that remains constant (and close to one) along this parameter slice ($\lambda_3 = W_{EE}(1 - \alpha)$) (Fig. 4B, bottom). Interestingly, we find that decay for the stationary autocovariance function for the inhibitory population (Fig. 4D and 4E

bottom; see Eq. 9 in Section 5) is well approximated by

$$\psi = \max\left(\mathrm{Re}(\lambda_1), \mathrm{Re}(\lambda_2)\right).$$

From this link, we see that when $|W_{EI}| = W_{IE}$ is large, then ψ is small, meaning the timescale of inhibition is fast. This allows the inhibitory population to rapidly and effectively cancel the net excitatory inputs (Fig. 4C, top). We observe that in this parameter regime, Δr_I remains small, while $\Delta r_{E_1} \approx -\Delta r_{E_2}$, leading to strong negative correlations between E_1 and E_2. As $|W_{EI}| = W_{IE}$ decreases, ψ increases towards one, which slows down the inhibitory timescale (Fig. 4D, top). This slower cancellation of the excitatory currents allows for larger deviations away from baseline for all neuronal populations. However, since the system is still stable, we observe that the populations co-vary together, leading to correlated excursions in the rates.

In total, the ISN regime yielded a more robust set of parameter values corresponding to high correlations across the segregated excitatory populations than the non-ISN regime observed previously. However, even in this improved scenario, the viable parameter regime is still limited to a relatively thin band, and further, this band lies precariously close to regions of instability.

2.6 Strong recurrent excitation with clustered inhibition

The fine tuning required to capture large $\mathrm{Corr}(E_1, E_2)$ despite having weak $E_1 \leftrightarrow E_2$ coupling ($\alpha << 1$) for both the purely excitatory and global inhibitory networks places doubt on these mechanisms being operative in real neuronal circuits. In this section we hypothesize that if the sources of inhibition for each excitatory subpopulation are similarly clustered, then this decoupling of inhibition may permit a larger stable region of positive correlations in the ISN regime, largely by limiting the effects of the anticorrelating $E_1 \to I \to E_2$ and $E_2 \to I \to E_1$ pathways.

We implemented inhibition to be co-clustered with the excitatory subpopulations by separating the inhibitory population in two, with I_1 and I_2 corresponding to the respective excitatory populations E_1 and E_2 (Fig. 5A). In this case, each E_i/I_i cluster constitutes an ISN ($i = 1, 2$). The model contains no inter-population connections except those between E_1 and E_2, and without any source of shared input correlations ($c = 0$). We have again assumed symmetry in the connection strengths such that the pairs (E_1, I_1) and (E_2, I_2) are identical in their connectivity and dynamics.

If we again fix W_{EE}, W_{II} and proceed by exploring the space of W_{EI}, W_{IE} connections, we find that this network structure now yields a robust region in which correlations are strong and positive (Fig. 5B, green). This can only be due to the strong dynamic recruitment of the inter-population connection αW_{EE}. This result emphasizes three important points. First, that there exists a large space of connection parameters in which our criteria (large $\mathrm{Corr}(E_1, E_2)$ and $\alpha << 1$) may be met. Given the heterogeneity of neural circuits and plasticity of connections within cortex this parametric result is much more satisfying than a fine-tuned solution like that required in the model with global inhibition (Sections 2.4, 2.5). Second, this result does not depend on the presence of external correlated fluctuations. Thirdly, this result is robust to the presence of external correlated input noise as it would only further amplify the observed correlations.

A natural question is whether incorporating inter-population $E_i \leftrightarrow I_j$ or $I_i \leftrightarrow I_j$ connections would affect this result. We therefore considered fixed values of W_{EI}, W_{IE}, and W_{II}, and introduced scaling parameters β, γ, ζ, respectively, to adjust the between-population strengths of each connection (Fig. 5C). We found that only $W_{II} > 0$ was able to further enhance correlations above the value we found when $\zeta = 0$ (Fig. 5C, left). In contrast, any non-zero values of β, γ only reduced correlations (Fig.

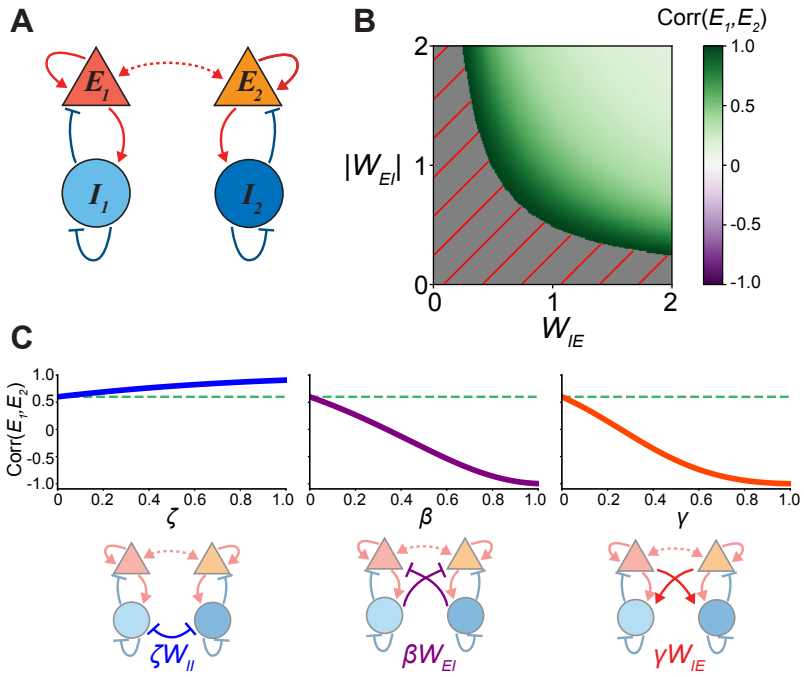


Figure 5. Segregated I subpopulations produce robust positive correlations. **A:** Model schematic. Input structure is consistent with Fig. 3A but omitted for clarity. **B:** $\text{Corr}(E_1, E_2)$ as a function of W_{EI}, W_{IE} with $c = 0$. **C:** $\text{Corr}(E_1, E_2)$ as a function of added connections between I_1, I_2 (left); $I_1 \to E_2$ and $I_2 \to E_1$ (middle); $E_1 \to I_2$ and $E_2 \to I_1$ (right). Added connections W_{ij} are initialized to the same as elsewhere in the network, and scaled by: ζ, $I \leftrightarrow I$; $\beta, I \to E$; $\gamma, E \to I$. Dashed turquoise line denotes $\zeta, \beta, \gamma = 0$, $W_{EI} = W_{IE} = 1$.

5C, middle and right). This is due to the same mechanism discussed previously in which strong excitatory recruitment of inhibition induces anti-correlations between the populations. Further, this same relationship also held when W_{EI}, W_{IE}, W_{II} were co-varied. Only $I \to I$ connections served as a correlating force; all others induced a reduction in correlations (AFig. 2). Hence, we conclude that while inhibition can be promiscuously connected with other inhibitory units, it must be strongly co-clustered with excitatory subpopulations and sparse in its connectivity with other excitatory subpopulations to yield the significant positive inter-population excitatory correlations observed in Kim et al. (2018b).

3 Discussion

In this study we sought to uncover possible neural circuit mechanisms underpinning the experimental observation that pyramidal neurons projecting to different downstream targets connect with a much lower probability than random pairs of excitatory neurons, yet still exhibit correlated variability that is almost as large as the rest of mouse V1 (Kim et al., 2018b). Notably, the magnitude of these correlations is much stronger than would be predicted given their weak connectivity. We found that a model with global inhibition resulted in highly constrained regions in which the data could be matched, encompassing two distinct solutions. In the case of weak network coupling, positive correlations resulted from two forms of *inheritance model*: either $I \to E$ connections induced increased correlated activity through I affecting both excitatory populations in

the same way, or an unobserved external source of strong correlations fed these fluctuations across both E units. When connectivity strengths grew, placing the circuit in an inhibition-stabilized (ISN) regime, the network needed to live right at the edge of stability to observe positive correlations. By contrast, we found that a more generally robust solution in the ISN regime could be achieved by splitting the inhibitory population into two separate subpopulations co-clustered with one of the excitatory subnetworks.

We argue that, on the basis of this robustness, our results thus predict that inhibition should cluster together with excitation in mouse sensory cortex with a specificity that mirrors that of the excitatory connectivity. The other inferred models by contrast depend upon narrow parameter regimes to capture experimental observations. This fragility would require significant constraints on the properties of neural circuits. Yet, connections are plastic, connection strengths are heterogeneous, and neuron properties are affected by neuromodulation (Turrigiano, 2008; Marder, 2012). Given this stochasticity in the circuit structure itself, a fine-tuned solution is unlikely to capture the data.

Rigorous experimental validation of our model predictions could be obtained through physiological or connectomics experiments which specifically target the relationship between excitatory projection neurons and local inhibitory neurons. While it is well-appreciated that inhibitory interneurons are very diverse in physiology and connectivity (Tremblay et al., 2016), we did not explicitly model this diversity in our study. Nevertheless, we anticipate that parvalbumin (PV)- positive cells may display the identified signatures of our I units, as they appear to play a critical role in stabilizing excitatory activity (Bos et al., 2020). Recent experimental evidence appears to support this claim from the perspective of stimulus tuning: while PV cells connect with most nearby pyramidal neurons, they were found to more strongly connect with those whose tuning properties they share (Znamenskiy et al., 2018).

Recent theoretical work has argued that E/PV assembly formation requires plasticity from both $E \rightarrow$ PV and PV $\rightarrow E$ connections (Mackwood et al., 2021). This bidirectionality could result in local, winner-take-all effects in $E \leftrightarrow I$ connectivity as any discrepancies in functional response properties between nearby pyramidal cells will bias the PV connectivity. This could result in the more specific co-clustering of inhibition we predict. Motivated by these results, a potential indirect way to differentiate between the global and clustered inhibition models would be to record activity of AL- and PM-projecting neurons together with inhibitory interneurons. Comparison of their respective tuning functions could suggest whether the inhibitory cell is biased in its connectivity (by extension of Znamenskiy et al. (2018)). Indeed, Najafi et al. (2020) recently argued for co-clustered excitation-inhibition in the context of posterior parietal cortex decision circuitry on the basis of neural response properties. Furthermore, in mouse visual cortex, it has been shown that PM and AL exhibit distinct functional representations with some overlap (Andermann et al., 2011), consistent with the tuning properties of V1 projection neurons (Kim et al., 2018b). Of course, it is possible that inhibitory-excitatory interactions may span a continuum between the global and clustered motifs identified here. This raises the possibility that heterogeneity in inhibitory connectivity motifs at small spatial scales may explain heterogeneity in pairwise covariance between AL- and PM-projecting pyramidal cells.

A central issue in the extension of our results concerns the dynamical regime of cortex, a topic which has received a significant amount of attention lately (Ahmadian and Miller, 2021; Morales et al., 2021; Huang, 2021). One question concerns whether intracortical interactions are strong enough to require inhibition as a key stabilizer of activity, that is, whether sensory cortex is an inhibition-stabilized network (Tsodyks et al., 1997; Sadeh and Clopath, 2021). Theoretical work predicts that in this regime

the ratio of excitatory to inhibitory input drive to a neuron decreases with increasing stimulus intensity (Rubin et al., 2015). Recent experimental evidence from recordings of mouse primary visual cortex supports this claim (Adesnik, 2017). Another study used optogenetic perturbation of inhibitory neurons across mouse cortex to test for inhibition-stabilization without sensory stimulation, finding evidence that all considered cortical regions operate as an ISN (Sanzeni et al., 2020).

Given this evidence for an ISN regime, a second question regards whether the network dynamics are poised near a change in stability. In our model, loss of stability would result in large positive correlations through a slowing down of the dynamics (Fig. 4). Analysis of large-scale recordings in mice has suggested that cortex may in fact live close to an instability (Morales et al., 2021). This could suggest that either the global or clustered inhibition model in an ISN regime may explain the data. Together with the foregoing evidence that PV and E neurons sharing tuning properties connect more strongly, we argue that this further supports a model of co-clustered inhibition.

Other mechanisms by which correlations can grow near a change in stability have been identified in previous studies. Ginzburg and Sompolinsky (1994) observed that near a bifurcation - in their case, a saddle node or Hopf - correlations in a weakly connected network grow from $\mathcal{O}(1/N)$ to near $\mathcal{O}(1)$ where N is the network size, together with a slowing down in the dynamics. Darshan et al. (2018) derived conditions on what they term the interaction matrix (similar to our $\mathbf{W}$ matrix) under which correlations are amplified without critical slowing down. These network models thus suggest distinct mechanisms by which our results could be extended to spatially-distributed spiking network models. Additionally, Litwin-Kumar and Doiron (2012) studied the effect of clustered connectivity in balanced spiking networks on the structure of correlations, however this work did not compare across-cluster to within-cluster correlations. Rosenbaum et al. (2017) did consider a structure similar to our three-population global inhibition motif, demonstrating that, consistent with our conclusions, a spatially distributed spiking neural network with distinct subpopulations would show close to zero correlations on average due to strong positive correlations within a cluster and large negative correlations between the two clusters. Yet it remains for future work to determine the precise parametric values to recapitulate our results in spiking neural network models.

Our work can be seen as a case study of a particular network structure in the context of the theoretical investigation of dynamics on graphs (that is, a collection of nodes and edges). In general, graphical analysis has been used in a wide range of neuroscientific applications, from the determination of fixed points of dynamics (Morrison and Curto, 2019) to network controllability (Kim et al., 2018a). In relating connectivity motifs (elements of $\mathbf{W}$ and their combinations) to correlation structure in the circuit, our approach relates to a more general mathematical concept of relating process motifs on networks to underlying structure motifs of the graph (Schwarze and Porter, 2021).

Ultimately, this work demonstrates how ostensibly straight-forward observations of connectivity and response properties from cortical cells have the capacity to lend fruitful insight into the structural and dynamical regimes of cortex, which are critical to further understanding of information processing in the brain.

4 Acknowledgments

A.N., M.P.G., and G.H. thank the Simons Foundation SCGB Undergraduate Research Fellowship (SURF) for fostering this collaboration. A.N. was funded by Simons Foundation SURF. G.H. was supported by the Burroughs Wellcome Fund's Career Award at the Scientific Interface. B.D. is supported by National Institutes of Health (NIH) (grant no. 1U19NS107613-01, R01EB026953), Vannevar Bush faculty fellowship

(no. N00014-18-1-2002) and the Simons Foundation Collaboration on the Global Brain.

5 Methods

5.1 Firing rate model

As done previously (Kanashiro et al., 2017; Getz et al., 2022), we consider the firing rate dynamics of neuronal populations A given by the following

$$\tau_A \frac{dr_A}{dt} = -r_A + f_A\left(\mu_A + \sum_B J_{AB} r_B + \hat{\sigma}_A\left[\sqrt{1-c}\cdot x_A(t) + \sqrt{c}\cdot x_s(t)\right]\right),$$

where τ_A is the time constant, μ_A is a constant stimulus drive, and J_{AB} is the strength of connections from population B to A. The stochastic processes $x_A(t)$ and $x_s(t)$ represent private and shared global fluctuations, respectively. Each is taken to be the limiting process from

$$\tau_x \frac{dx}{dt} = -x + \sqrt{\tau_x}\xi_x(t),$$

for $\tau_x \to 0$, with $\langle \xi_i(t)\rangle = 0$ and $\langle \xi_i(t)\xi_i(t')\rangle = \delta(t-t')$. Intuitively, one may think of $x(t)$ as a "smoothed" white noise process (Kanashiro et al., 2017). The parameter $c \in [0,1]$ scales the proportion of shared noise relative to private noise, while $\hat{\sigma}_A$ represents the total intensity of the fluctuations.

We assume that the system of equations has an equilibrium point at r_{ss}, and that the noise is weak enough so that the fluctuations about this equilibrium ($\Delta r := r - r_{\text{ss}}$) can be approximated by

$$\tau_A \frac{d\Delta r_A}{dt} = -\Delta r_A + L_A \sum_B J_{AB}\Delta r_B + L_A\hat{\sigma}_A\left[\sqrt{1-c}\cdot x_A(t) + \sqrt{c}\cdot x_s(t)\right],$$

where $L_A = f_A'(r_{\text{ss}})$ is the gain of population A at the equilibrium point. We define the effective coupling as $W_{AB} := L_A J_{AB}$ and $\sigma_A := L_a\hat{\sigma}_A$, and approximate $x_A(t)$ and $x_s(t)$ as independent, zero-mean Gaussian processes $\xi_A(t)$ and $\xi_s(t)$ satisfying $\langle \xi(t)\xi(t')\rangle = \delta(t-t')$. This yields Eq. 1, which in matrix form can be written as

$$\mathbf{T}\frac{d\boldsymbol{\Delta r}}{dt} = (\mathbf{W} - \mathbf{I})\Delta\mathbf{r}(t) + \mathbf{D}\boldsymbol{\xi}(t). \tag{8}$$

For notational simplicity, throughout we will assume unit time constants $\tau_A = 1$, so that $\mathbf{T} = \mathbf{I}$. For example, in the case of two excitatory populations and one inhibitory population $\{E_1, E_2, I\}$ the matrices are

$$\mathbf{W} = \begin{bmatrix} W_{E_1E_1} & W_{E_1E_2} & W_{E_1I} \\ W_{E_2E_1} & W_{E_2E_2} & W_{E_2I} \\ W_{IE_1} & W_{IE_2} & W_{II} \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} \sqrt{(1-c)}\cdot\sigma_{E_1} & 0 & 0 & \sqrt{c}\cdot\sigma_{E_1} \\ 0 & \sqrt{(1-c)}\cdot\sigma_{E_2} & 0 & \sqrt{c}\cdot\sigma_{E_2} \\ 0 & 0 & \sigma_I & 0 \end{bmatrix}.$$

The network structure is determined through the weight matrix $\mathbf{W}$. Since we are explicitly interested in segregated excitatory populations, we consider weak the cross-population connections and set

$$W_{E_2E_1} = \alpha W_{E_1E_1}, \qquad W_{E_1E_2} = \alpha W_{E_2E_2}$$

for $\alpha \in (0,1)$. The two excitatory populations, E_1 and E_2, are increasingly disconnected as $\alpha \to 0$. To obtain analytical expressions and constrain the searchable parameter space, we assume various symmetries in the network connectivity. Specifically, we consider the following forms for connectivity matrices for the two (Fig. 1A), three (Fig. 3A) and four (Fig. 4A) population models:

$$\mathbf{W} = \begin{bmatrix} W_{EE} & \alpha W_{EE} \\ \alpha W_{EE} & W_{EE} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} W_{EE} & \alpha W_{EE} & W_{EI} \\ \alpha W_{EE} & W_{EE} & W_{EI} \\ W_{IE} & W_{IE} & W_{II} \end{bmatrix},$$

and

$$\mathbf{W} = \begin{bmatrix} W_{EE} & \alpha W_{EE} & W_{EI} & \beta W_{EI} \\ \alpha W_{EE} & W_{EE} & \beta W_{EI} & W_{EI} \\ W_{IE} & \gamma W_{IE} & W_{II} & \zeta W_{II} \\ \gamma W_{IE} & W_{IE} & \zeta W_{II} & W_{II} \end{bmatrix},$$

where $\beta, \gamma, \zeta \in (0,1)$.

5.2 Covariance calculation

The autocovariance function for the OU process defined in Eq. 8 is given by

$$\widetilde{\mathbf{C}}(h) = \langle \Delta\mathbf{r}(t)\Delta\mathbf{r}(t+h)\rangle.$$

Let $\mathbf{M} = \mathbf{W} - \mathbf{I}$ and define $\mathbf{\Sigma} := \widetilde{\mathbf{C}}(0) = \langle \Delta\mathbf{r}(t)\Delta\mathbf{r}(t)\rangle$ as the stationary covariance matrix. Then $\mathbf{\Sigma}$ is obtained as the solution to the *Lyapunov equation* $-\mathbf{M\Sigma} + \mathbf{\Sigma}(-\mathbf{M})^\top = \mathbf{D}\mathbf{D}^\top$ (Gardiner, 2009). It follows that

$$\widetilde{\mathbf{C}}(h) = \begin{cases} e^{-\mathbf{M}h} \cdot \mathbf{\Sigma}, & h < 0 \\ \mathbf{\Sigma} \cdot e^{\mathbf{M}^\top h}, & h \geq 0. \end{cases} \tag{9}$$

Integrating $\widetilde{\mathbf{C}}(h)$ in each element over long times h yields the following compressed form for the *long-time covariance matrix* $\mathbf{C}$

$$\begin{aligned} \mathbf{C} &= \int_{-\infty}^{\infty} \widetilde{\mathbf{C}}(h)\, dh \\ &= \mathbf{M}^{-1}\mathbf{D}\left(\mathbf{M}^{-1}\mathbf{D}\right)^\top. \end{aligned}$$

If $\mathbf{C}_V = \sqrt{\text{diag}(\mathbf{C})}$, then the correlation matrix is obtained

$$\boldsymbol{\rho} = \mathbf{C}_V^{-1}\mathbf{C}\mathbf{C}_V^{-1}. \tag{10}$$

5.3 Path expansion

If the spectral radius $s(\mathbf{W}) = \max\{|\lambda_i| : \lambda_i \text{ is an eigenvalue of } \mathbf{W}\} < 1$, then $\mathbf{M}^{-1}$ has a convergent series representation

$$-\mathbf{M}^{-1} = (\mathbf{I} - \mathbf{W})^{-1} = \sum_{k=0}^{\infty} \mathbf{W}^k$$

known as a *Neumann series* (Einsiedler et al., 2017). Intuitively, one may think of the Neumann series as a matrix analogue of the familiar geometric series. Under this representation, the long-time covariance matrix is

$$\mathbf{C} = \left(\sum_{k=0}^{\infty} \mathbf{W}^j\right) \mathbf{D}\mathbf{D}^\top \left(\sum_{k=0}^{\infty} \mathbf{W}^k\right)^\top.$$

It is useful to rewrite this expansion as

$$\mathbf{C} = \sum_{n=0}^{\infty} \left[\sum_{i=0}^{n} \mathbf{W}^{n-i} \mathbf{D}\mathbf{D}^\top \left(\mathbf{W}^\top\right)^i\right],$$

where the terms in the inner sum can be interpreted as contributions due to n^{th}-order paths through the network (Trousdale et al., 2012; Pernice et al., 2011).

If the outer sum converges quickly, the covariance matrix can be approximated as

$$\mathbf{C} \approx \sum_{n=0}^{N} \left[\sum_{i=0}^{n} \mathbf{W}^{n-i} \mathbf{D}\mathbf{D}^\top \left(\mathbf{W}^\top\right)^i\right].$$

The rate of convergence of this approximation depends on the magnitude of $s(\mathbf{W})$. In particular, the closer $s(\mathbf{W})$ is to 0, the faster the terms shrink. Consider the N-th order terms of this approximation,

$$\sum_{i=0}^{N} \mathbf{W}^{N-i} \mathbf{D}\mathbf{D}^\top \left(\mathbf{W}^\top\right)^i.$$

If $||\cdot||$ is the operator norm, then

$$\begin{aligned}\left\|\sum_{i=0}^{N} \mathbf{W}^{N-i} \mathbf{D}\mathbf{D}^\top \left(\mathbf{W}^\top\right)^i\right\| &\leq \sum_{i=0}^{N} \left\|\mathbf{W}^{N-i} \mathbf{D}\mathbf{D}^\top \left(\mathbf{W}^\top\right)^i\right\| \\ &\leq \sum_{i=0}^{N} \left\|\mathbf{D}\mathbf{D}^\top\right\| \cdot \left\|\mathbf{W}^{N-i} \left(\mathbf{W}^\top\right)^i\right\|\end{aligned}$$

Diagonalize $\mathbf{W}$ and write $\mathbf{W} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{W}$. It follows that

$$\begin{aligned}\left\|\mathbf{W}^{N-i} \left(\mathbf{W}^\top\right)^i\right\| &= \left\|\left(\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}\right)^{N-i} \left((\mathbf{P}^{-1})^\top \mathbf{\Lambda} \mathbf{P}^\top\right)^i\right\| \\ &= \left\|\mathbf{P}\mathbf{\Lambda}^{N-i}\mathbf{P}^{-1}(\mathbf{P}^{-1})^\top \mathbf{\Lambda}^i \boldsymbol{P}^\top\right\| \\ &\leq \|\mathbf{P}\|^2 \left\|\mathbf{P}^{-1}\right\|^2 \|\mathbf{\Lambda}\|^N\end{aligned}$$

Then

$$\begin{aligned}\sum_{i=0}^{N} \left\|\mathbf{D}\mathbf{D}^\top\right\| \cdot \left\|\mathbf{W}^{N-i} \left(\mathbf{W}^\top\right)^i\right\| &\leq N \left\|\mathbf{D}\mathbf{D}^\top\right\| \cdot \|\mathbf{P}\|^2 \cdot \left\|\mathbf{P}^{-1}\right\|^2 \cdot \|\mathbf{\Lambda}\|^N \\ &\leq N \left\|\mathbf{D}\mathbf{D}^\top\right\| \cdot \|\mathbf{P}\|^2 \cdot \left\|\mathbf{P}^{-1}\right\|^2 \cdot s(\mathbf{W})^N\end{aligned}$$

This bound shrinks quickly as $N \to \infty$ if $s(\mathbf{W})$ is small ($\ll 1$), as is the case when the system is in the weakly coupled regime.

5.3.1 Path expansion for weakly coupled $E_1 \leftrightarrow E_2$

In Fig. 2B we illustrate this quick convergence by showing the first three terms of this sum, namely

$$\begin{aligned} 0^{\text{th}} \text{ - order: } & \mathbf{D}\mathbf{D}^\top, \\ 1^{\text{st}} \text{ - order: } & \mathbf{W}\mathbf{D}\mathbf{D}^\top + \mathbf{D}\mathbf{D}^\top\mathbf{W}^\top, \\ 2^{\text{nd}} \text{ - order: } & \mathbf{W}^2\mathbf{D}\mathbf{D}^\top + \mathbf{W}\mathbf{D}\mathbf{D}^\top\mathbf{W}^\top + \mathbf{D}\mathbf{D}^\top\left(\mathbf{W}^\top\right)^2. \end{aligned}$$

Using these terms, the cross population covariance can be approximated by Eqn. 5 in the main text.

We note that for a n^{th}-order path, we multiply on the left and right by $\mathbf{C}_V^{-1}$ to obtain path contributions to the correlation matrix. In particular, we are interested in the contributions to $\rho_{E_1E_2}$ (that is, the element $\boldsymbol{\rho}_{1,2}$ of Eqn. 10).

5.4 Parameters & Simulations

All relevant code will be made available at the author's github upon publication. Simulations were performed using an Euler-Maruyama scheme with time constants $\tau_E = \tau_I = 15$ msec, $dt = 0.01$ msec.

Table 1. Strength of connections from pop. B (columns) to A (rows) for the weakly (strongly) coupled model.

W_{AB}	E	I
E	0.5 (1.15)	0.5 (0.8)
I	0.5 (0.8)	0.5 (0.5)

Table 2. Default parameter values. Changes to any parameter are indicated in the figure caption.

Parameter	Default value	Description
α	0.15	Inter-excitatory population strength
σ_A	1	Total intensity of outside fluctuations
c	0	Scales the proportion of shared noise relative to private noise

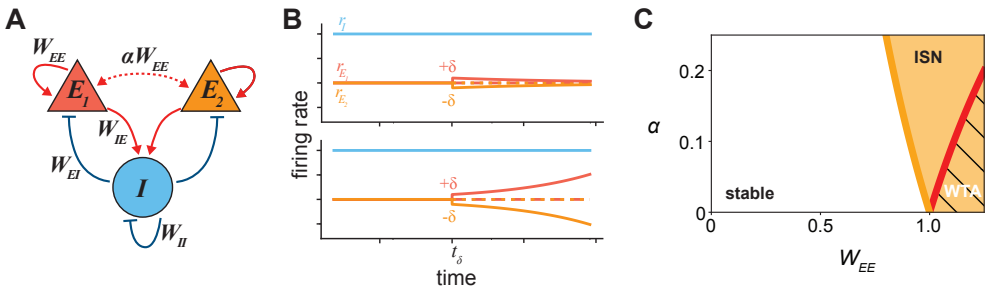
6 Supplemental Information

6.1 Network stability

The deterministic version of Eqn. 8 (Fig. 1A)

$$\frac{d\boldsymbol{\Delta}\mathbf{r}}{dt} = (\mathbf{W} - \mathbf{I})\boldsymbol{\Delta}\mathbf{r}(t)$$

is *asymptotically stable* if the eigenvalues λ_i of $\mathbf{W} - \mathbf{I}$ satisfy $\Re[\lambda_i] < 0$, meaning that a perturbation of the excitatory rates is quenched and rates are returned to their steady-state values (AFig. 1B, top) (Wiggins, 2003). An equivalent condition for stability is if the eigenvalues λ_i of $\mathbf{W}$ satisfy $\Re\{\lambda_i\} < 1$. We say that a network is stable if it admits a stable equilibrium solution, otherwise we say the network is unstable.



Supplemental Figure 1. Dynamical regimes and limitations on α. A: Network schematic. **B:** Illustrations of a small change in the input $+\delta$ to E_1 and $-\delta$ to E_2. Top: stable network regime; bottom: unstable (winner-take-all) regime. **C:** $W_{EE} - \alpha$ space. Yellow region: inhibition-stabilized (ISN); black hatched region: winner-take-all (unstable). Solid yellow line: $W_{EE} = 1/(1+\alpha)$. Solid red line: $\alpha = 1 - 1/W_{EE}$. Parameters as in Fig. 5. α changed to 0.1 for the unstable regime in B.

6.1.1 The inhibition-stabilized network (ISN)

A linear network is an *inhibition stabilized network* (ISN) (Ozeki et al., 2009) if it satisfies two conditions:

(1) The network is unstable in the absence of (dynamic) feedback inhibition,

(2) The network is stable with sufficiently strong inhibition.

We consider the conditions under which the global inhibition motif (i.e., two excitatory populations with one shared inhibitory population) is an ISN. The corresponding weight matrix is

$$\mathbf{W} = \begin{bmatrix} W_{EE} & \alpha W_{EE} & W_{EI} \\ \alpha W_{EE} & W_{EE} & W_{EI} \\ W_{IE} & W_{IE} & W_{II} \end{bmatrix},$$

which has eigenvalues

$$\lambda_1 = (1-\alpha)\cdot W_{EE},$$
$$\lambda_{2,3} = \frac{1}{2}\left[(1+\alpha)\cdot W_{EE} + W_{II} \pm \sqrt{(1+\alpha)^2 W_{EE}^2 + 8W_{EI}W_{IE} - 2(1+\alpha)W_{EE}W_{II} + W_{II}^2}\right].$$

We note that λ_1 does not depend on any of the inhibitory connections. As a result, if $\lambda_1 = (1-\alpha)\cdot W_{EE} > 1$ the system is unstable and inhibition is unable to stabilize it, so we necessarily require $(1-\alpha)\cdot W_{EE} < 1$. On the other hand, $\lambda_{2,3}$ do depend on the inhibitory connections. Absent feedback inhibition (i.e., $W_{EI} = 0$) these eigenvalues become

$$\lambda_2 = (1+\alpha)\cdot W_{EE} \text{ and } \lambda_3 = W_{II}.$$

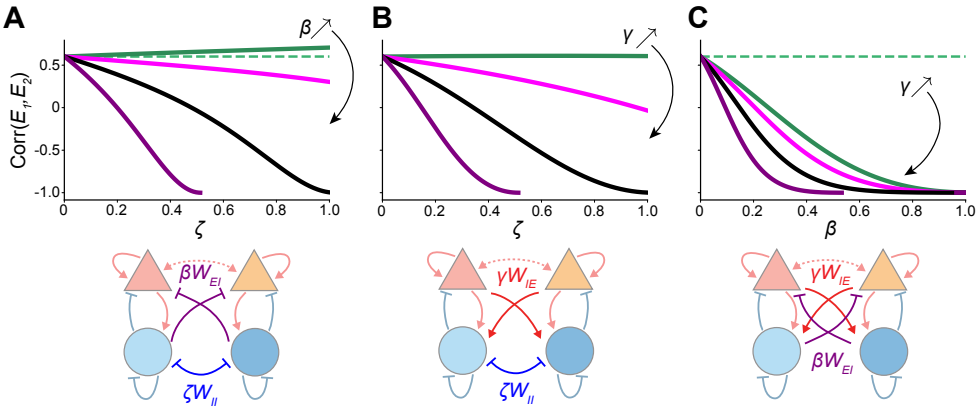
In this work the latter is always less than 1. Meanwhile, it is possible to increase recurrent excitation such that $\lambda_2 = (1+\alpha)\cdot W_{EE} > 1$. Unlike the previous condition derived with λ_1, we can choose inhibitory parameters W_{EI}, W_{II} such that this eigenvalue decreases below 1, restoring the stability of the system. Thus, this system lies in the ISN regime when

$$(1+\alpha)\cdot W_{EE} > 1 \text{ and } (1-\alpha)\cdot W_{EE} < 1.$$

If the first condition is satisfied, while the second condition is violated, the system exhibits winner-take-all dynamics, where one excitatory population increases away from steady state while the second decreases away from it (AFig. 1B, bottom). All three regions (non-ISN, ISN, and winner-take-all) are shown in AFig. 1C. The same constraint to lie in the ISN regime can also be derived for the specific inhibition motif.

6.2 Covarying cross-population connections

It is possible that in the segregated E_i/I_i subpopulation model (Fig. 5A), covarying cross-population connections might induce synergetic effects different from those observed by adding singular bidirectional connections (Fig. 5C). We tested this numerically by adding pairwise combinations of $E \to I$, $I \to E$, and $I \to I$ (AFig. 2A-C). Only when ζ (scaling of $I \to I$) dominated either β or γ was an increase in correlations observed; $E \to I$ and $I \to E$ always reduced correlations, consistent with the results presented above.



Supplemental Figure 2. Covarying cross-population connections in segregated ISN. A: Top: $Corr(E_1, E_2)$ as a function of ζ; colored lines indicate different values of β. Black lines indicate when $\zeta = \beta$. Bottom: network schematics indicating connection weights co-varied in the above plot. Green dashed line indicates value of $Corr(E_1, E_2)$ for $W_{EI} = W_{IE}$. Parameters as in Fig. 5. **B:** same as (A) for ζ and γ. **C:** same as (A) for β and γ.

References

H. Adesnik. Synaptic mechanisms of feature coding in the visual cortex of awake mice. *Neuron*, 95(5):1147–1159, 2017.

Y. Ahmadian and K. D. Miller. What is the dynamical regime of cerebral cortex? *Neuron*, 109(21):3373–3391, 2021.

M. L. Andermann, A. M. Kerlin, D. K. Roumis, L. L. Glickfeld, and R. C. Reid. Functional specialization of mouse higher visual cortical areas. *Neuron*, 72(6): 1025–1039, 2011.

T. Biswas and J. E. Fitzgerald. Geometric framework to predict structure from function in neural networks. *Physical Review Research*, 4(2):023255, 2022.

H. Bos, A.-M. Oswald, and B. Doiron. Untangling stability and gain modulation in cortical circuits with multiple interneuron classes. *bioRxiv*, pages 2020–06, 2020.

M. R. Cohen and A. Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811–819, 2011.

L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015.

R. Darshan, C. Van Vreeswijk, and D. Hansel. Strength of correlations in strongly recurrent neuronal networks. *Physical Review X*, 8(3):031072, 2018.

A. Das and I. R. Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, 23(10):1286–1296, 2020.

B. Doiron, A. Litwin-Kumar, R. Rosenbaum, G. K. Ocker, and K. Josić. The mechanics of state-dependent neural correlations. *Nature neuroscience*, 19(3):383–393, 2016.

M. Einsiedler, T. Ward, et al. *Functional analysis, spectral theory, and applications*, volume 104. Springer, 2017.

A. A. Faisal, L. P. Selen, and D. M. Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292–303, 2008.

C. Gardiner. *Stochastic methods*, volume 4. Springer Berlin, 2009.

M. P. Getz, C. Huang, and B. Doiron. Subpopulation codes permit information modulation across cortical states. *bioRxiv*, pages 2022–09, 2022.

I. Ginzburg and H. Sompolinsky. Theory of correlations in stochastic neural networks. *Physical review E*, 50(4):3171, 1994.

K. M. Hagihara, A. W. Ishikawa, Y. Yoshimura, Y. Tagawa, and K. Ohki. Long-range interhemispheric projection neurons show biased response properties and fine-scale local subnetworks in mouse visual cortex. *Cerebral Cortex*, 31(2):1307–1315, 2021.

S. B. Hofer, H. Ko, B. Pichler, J. Vogelstein, H. Ros, H. Zeng, E. Lein, N. A. Lesica, and T. D. Mrsic-Flogel. Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nature neuroscience*, 14(8):1045–1052, 2011.

C. Huang. Modulation of the dynamical state in cortical network models. *Current opinion in neurobiology*, 70:43–50, 2021.

T. Kanashiro, G. K. Ocker, M. R. Cohen, and B. Doiron. Attentional modulation of neuronal variability in circuit models of cortex. *Elife*, 6:e23978, 2017.

J. Z. Kim, J. M. Soffer, A. E. Kahn, J. M. Vettel, F. Pasqualetti, and D. S. Bassett. Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nature physics*, 14(1):91–98, 2018a.

M.-H. Kim, P. Znamenskiy, M. F. Iacaruso, and T. D. Mrsic-Flogel. Segregated subnetworks of intracortical projection neurons in primary visual cortex. *Neuron*, 100 (6):1313–1321, 2018b.

H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011.

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

O. Mackwood, L. B. Naumann, and H. Sprekeler. Learning excitatory-inhibitory neuronal assemblies in recurrent networks. *Elife*, 10:e59715, 2021.

E. Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1): 1–11, 2012.

Y. Mishchencko, J. T. Vogelstein, and L. Paninski. A bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *The Annals of Applied Statistics*, pages 1229–1261, 2011.

G. B. Morales, S. Di Santo, and M. A. Muñoz. Quasi-universal scaling in mouse-brain neuronal activity stems from edge-of-instability critical dynamics. *arXiv preprint arXiv:2111.12067*, 2021.

K. Morrison and C. Curto. Predicting neural network dynamics via graphical analysis. In *Algebraic and Combinatorial Computational Biology*, pages 241–277. Elsevier, 2019.

F. Najafi, G. F. Elsayed, R. Cao, E. Pnevmatikakis, P. E. Latham, J. P. Cunningham, and A. K. Churchland. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron*, 105(1): 165–179, 2020.

G. K. Ocker, Y. Hu, M. A. Buice, B. Doiron, K. Josić, R. Rosenbaum, and E. Shea-Brown. From the statistics of connectivity to the statistics of spike times in neuronal networks. *Current opinion in neurobiology*, 46:109–119, 2017.

H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62 (4):578–592, 2009.

A. M. Packer and R. Yuste. Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *Journal of Neuroscience*, 31(37):13260–13271, 2011.

V. Pernice, B. Staude, S. Cardanobile, and S. Rotter. How structure determines correlations in neuronal networks. *PLoS computational biology*, 7(5):e1002059, 2011.

A. Renart, N. Brunel, and X.-J. 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.

A. Renart, J. De La Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris. The asynchronous state in cortical circuits. *science*, 327(5965):587–590, 2010.

R. Rosenbaum, M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron. The spatial structure of correlated neuronal variability. *Nature neuroscience*, 20(1):107–114, 2017.

Y. Roudi, B. Dunn, and J. Hertz. Multi-neuronal activity and functional connectivity in cell assemblies. *Current opinion in neurobiology*, 32:38–44, 2015.

D. B. Rubin, S. D. Van Hooser, and K. D. Miller. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015.

S. Sadeh and C. Clopath. Inhibitory stabilization and cortical computation. *Nature Reviews Neuroscience*, 22(1):21–37, 2021.

A. Sanzeni, B. Akitake, H. C. Goldbach, C. E. Leedy, N. Brunel, and M. H. Histed. Inhibition stabilization is a widespread property of cortical networks. *Elife*, 9:e54875, 2020.

A. C. Schwarze and M. A. Porter. Motifs for processes on networks. *SIAM Journal on Applied Dynamical Systems*, 20(4):2516–2557, 2021.

R. Tremblay, S. Lee, and B. Rudy. Gabaergic interneurons in the neocortex: from cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

J. Trousdale, Y. Hu, E. Shea-Brown, and K. Josić. Impact of network structure and cellular response on spike time correlations. *PLoS computational biology*, 8(3): e1002408, 2012.

M. V. Tsodyks, W. E. Skaggs, T. J. Sejnowski, and B. L. McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *Journal of neuroscience*, 17 (11):4382–4388, 1997.

G. G. Turrigiano. The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435, 2008.

A. E. Urai, B. Doiron, A. M. Leifer, and A. K. Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature neuroscience*, 25 (1):11–19, 2022.

S. Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*, volume 2. Springer, 2003.

P. Znamenskiy, M.-H. Kim, D. R. Muir, M. F. Iacaruso, S. B. Hofer, and T. D. Mrsic-Flogel. Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *Biorxiv*, page 294835, 2018.