Exact linear theory of perturbation response in a space- and feature-dependent cortical circuit model

Ho Yin Chau[a,1], **Kenneth D. Miller**[a,b,*], **and Agostina Palmigiano**[a,c,*]

This manuscript was compiled on January 24, 2025

What are the principles that govern the responses of cortical networks to their inputs and the emergence of these responses from recurrent connectivity? Recent experiments have probed these questions by measuring cortical responses to two-photon optogenetic perturbations of single cells in the mouse primary visual cortex. A robust theoretical framework is needed to determine the implications of these responses for cortical recurrence. Here we propose a novel analytical approach: a formulation of the dependence of cell-type-specific connectivity on spatial distance that yields an exact solution for the linear perturbation response of a model with multiple cell types and space- and feature-dependent connectivity. Importantly and unlike previous approaches, the solution is valid in regimes of strong as well as weak intra-cortical coupling. Analysis reveals the structure of connectivity implied by various features of single-cell perturbation responses, such as the surprisingly narrow spatial radius of nearby excitation beyond which inhibition dominates, the number of transitions between mean excitation and inhibition thereafter, and the dependence of these responses on feature preferences. Comparison of these results to existing optogenetic perturbation data yields constraints on cell-type-specific connection strengths and their tuning dependence. Finally, we provide experimental predictions regarding the response of inhibitory neurons to single-cell perturbations and the modulation of perturbation response by neuronal gain; the latter can explain observed differences in the feature-tuning of perturbation responses in the presence vs. absence of visual stimuli.

Recurrent neural networks | Optogenetic perturbation | Mouse Primary Visual Cortex

Significance Statement

The cerebral cortex is strongly recurrently connected with complex wiring rules. This circuitry can now be probed by studying responses to optogenetic perturbations of one or small numbers of cells. However, we currently lack a general theory connecting these responses to underlying circuitry. Here we develop a novel, exactly solvable theory to determine responses to small perturbations from the underlying connectivity. Analysis of these equations reveals simple rules that govern perturbation response patterns. Comparison with experimental data yields new constraints on the connectivity parameters. The theory yields predictions for the responses of unmeasured cell types and in new experimental conditions.

In recent years there have been a number of experiments utilizing holographic perturbation techniques to probe recurrent neuronal circuitry. In layers 2/3 (L2/3) of the mouse primary visual cortex (V1), such experiments have revealed complex rules governing the perturbation response of neurons that depend on the spatial locations and orientation tunings of both the perturbed and the unperturbed neurons (1–7).

A common approach to making sense of this rich structure is to model mouse V1 L2/3 with a linear, recurrently-connected firing rate model where connectivity strength depends on the spatial location, orientation tuning, and cell type of the pre- and post-synaptic neurons (2). While such models provide much simpler descriptions than biophysical spiking models and are analytically tractable for weak connectivity (spectral radius of weight matrix < 1), there is still a lack of a more general understanding of how the perturbation response is related to the underlying connectivity structure.

Here we introduce a novel analytical approach to the problem. First, we show that an exponential-like spatial connectivity kernel is a good descriptor of the product of connection probability and synaptic strength. This choice of kernel allows us to derive an exact solution for the linear perturbation response of recurrently connected networks with multiple cell types that is valid regardless of the spectral radius of the weight matrix. As this formulation holds for any circuit coupling strength, it allows one to investigate perturbation responses of inhibition stabilized networks (ISNs) (8, 9), which appear to describe cortical circuits (10) and which may be characterized by large negative eigenvalues.

The general solution for the circuit involving an arbitrary number of cell-types and connectivity length scales is complex, and does not easily provide intuitive insight. However, for the special case of an excitatory/inhibitory (E-I) network in which connectivity width depends only on presynaptic cell types, we discover simple mathematical rules that govern the relationship between connectivity structure and single-cell perturbation response. These insights allow us to infer various

Author affiliations: [a]Center for Theoretical Neuroscience, College of Physicians and Surgeons and Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY; [b]Dept. of Neuroscience, Swartz Program in Theoretical Neuroscience, Kavli Institute for Brain Science, College of Physicians and Surgeons and Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York; [c]Gatsby Computational Neuroscience Unit, University College London; [*]These authors contributed equally to this work

H.Y.C., K.D.M. and A.P. designed research. H.Y.C. performed research and contributed analytic tools. H.Y.C., K.D.M., and A.P. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: hc3190@cumc.columbia.edu

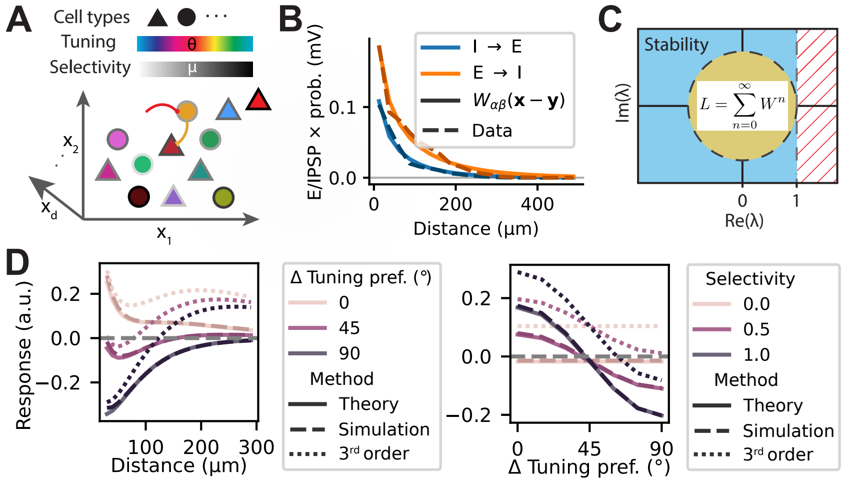

Fig. 1. Exact linear response theory vs simulation and common approximations. A) Schematic of model. Neurons are located in a d-dimensional space with N_c cell types, and have feature tuning preferences $\theta \in [-\pi, \pi)$, and feature selectivities $\mu \in [0, 1]$. **B)** Connectivity function in a simplified 2D model, $W_{\alpha\beta}(\boldsymbol{x} - \boldsymbol{y})$ (equation 3), fitted to the product of connection probability (11) and connection strength (12) between excitatory and inhibitory neurons. **C)** Region of convergence (yellow, all eigenvalues λ of the weight matrix satisfying $|\lambda| < 1$) for the matrix inverse expansion used in existing theoretical analyses of perturbation response, compared to the region of stability (blue, all $\mathrm{Re}(\lambda) < 1$, assuming cell-type-independent time constants), for which our theory applies. **D)** Comparison between theory, simulation, and 3rd order matrix inverse expansion for the single-cell perturbation response of an E-I model with spectral radius of 1.8. Left: Response of excitatory neurons as a function of distance to the perturbed excitatory neuron, for different feature tuning preferences. Right: Response of excitatory neurons as a function of difference in feature tuning preference from the perturbed excitatory neuron, for different feature selectivities.

constraints that the cortical connectivity should satisfy in order to explain existing optogenetic perturbation data.

We break down our analysis in four sections: first, we examine the condition for the circuit to exhibit mean suppression in response to the perturbation of an excitatory neuron, as observed in (1). This is followed by two sections on the analysis of distance and feature preference dependence of the perturbation response respectively. In particular, in the first, we characterize the number and location of spatial zero crossings of the network response (*i.e.* transitions between mean excitatory and mean inhibitory response with distance from the perturbation). Finally we study the joint dependence of the perturbation response on distance and feature tuning, specifically the relationship between feature-specific amplification/suppression and distance.

To validate our theoretical findings, we establish several predictions that can be tested experimentally. First, since existing perturbation data mainly studies the response of excitatory neurons to excitatory neuron perturbation, we predict that the response of inhibitory neurons should exhibit less suppression and a broader spatial profile than excitatory neurons. Second, since the perturbation experiments may be performed with or without the simultaneous presentation of visual stimuli, we predict that the absence of visual stimuli, which reduces firing rate and hence neuronal gain, may result in feature tuning dependence of perturbation response which is opposite to that when visual stimuli are present. Finally, we predict that the absence of visual stimuli should generally result in responses with less suppression and a broader Mexican-hat profile response, possibly eliminating the presence of zero-crossings altogether.

Results

We study responses to moderate single-cell perturbations. Because these perturbations are small, we expect a linear theory to be adequate. To this end, we consider a linear recurrent neuronal network with N_c cell types and d spatial dimensions (Figure 1A). Each neuron is uniquely indexed by the four-tuple $(\alpha, \mu, \boldsymbol{x}, \theta) \in \mathbb{Z}_{N_c} \times [0, 1] \times \mathbb{R}^d \times \mathbb{S}^1$, representing cell type, feature selectivity, spatial location, and feature tuning preference respectively. The firing rate of neuron $(\alpha, \mu, \boldsymbol{x}, \theta)$ at time t is written $r_\alpha(\mu, \boldsymbol{x}, \theta, t)$, while the connectivity weight between postsynaptic neuron $(\alpha, \mu, \boldsymbol{x}, \theta)$ and presynaptic neuron $(\beta, \nu, \boldsymbol{y}, \phi)$ is denoted $W_{\alpha\beta}(\mu, \nu, \boldsymbol{x} - \boldsymbol{y}, \theta - \phi)$. Feature selectivity (*i.e.* how well tuned a neuron is) is assigned independently to each neuron and may be arbitrarily distributed with density $P_\alpha(\mu)$. The external input to each neuron is denoted $h_\alpha(\mu, \boldsymbol{x}, \theta)$. For the single-cell perturbations we are considering, h is a delta function given by equation 10. Taking the continuum limit for our analytical work, the dynamical equation of the network is given by equation 11. We are primarily interested in the steady-state response $r_\alpha(\mu, \boldsymbol{x}, \theta) = \lim_{t\to\infty} r_\alpha(\mu, \boldsymbol{x}, \theta, t)$, which exists if and only if the network is stable and is given by

$$r_\alpha(\mu, \boldsymbol{x}, \theta) = \sum_{\beta=0}^{N_c-1} \int_0^1 \int_{\mathbb{R}^d} \int_{-\pi}^{\pi} W_{\alpha\beta}(\mu, \nu, \boldsymbol{x} - \boldsymbol{y}, \theta - \phi) \, r_\beta(\nu, \boldsymbol{y}, \phi) P_\beta(\nu)\, d\phi d\boldsymbol{y} d\nu + h_\alpha(\mu, \boldsymbol{x}, \theta) \qquad [1]$$

In general, there is no closed-form analytical solution for arbitrary choices of W. Our key insight is that W can be chosen such that it captures the spatial dependence of the product of the connection probability and the synaptic strength between cells (Figure 1B), and admits a closed-form analytical solution, as we now explain.

We will make the common assumption that the dependence of W on space and feature can be factorized. The spatial dependence is commonly modeled as a Gaussian kernel (13–19), in accordance with the approximately Gaussian spatial profile of connection probability measured in mouse V1 L2/3 (11, 20). However, this choice of spatial kernel neglects the spatial decay of synaptic strength (12) and does not admit a closed-form solution for equation 1. Instead, we propose setting the spatial kernel as $G_d(r; \sigma^{-2})$, where r is the spatial distance, σ is the connectivity length scale, and $G_d(\|\cdot\|; \sigma^{-2})$ is the Green's function (effectively, the inverse) of the operator $\sigma^{-2} - \nabla^2$ in d-dimensions. Specifically, G_d is a monotonic, exponentially-decaying kernel given by

$$G_d(r; \lambda) = \frac{1}{(2\pi)^{\frac{d}{2}}} \left(\frac{\sqrt{\lambda}}{r} \right)^{\nu} K_\nu(\sqrt{\lambda} r) \qquad [2]$$

where $\nu = \frac{d}{2} - 1$ and $K_\nu(z)$ is the modified Bessel function of the second kind with order ν (SI section 1). In 1 and

3 dimensions, $G_d(r;\lambda)$ is proportional to $e^{-\sqrt{\lambda}r}$ and $\frac{e^{-\sqrt{\lambda}r}}{r}$ respectively. We combine data from (11) and (12) to compute the product of connection probability and connection strength as a function of distance between excitatory and inhibitory neurons in mouse V1 L2/3. We find that our kernel can exactly capture this dependence (Figure 1B; Materials and Methods), with best-fit E → I and I → E connectivity widths given by $\sigma_{\mathrm{E}} = (150 \pm 11)\,\mu\mathrm{m}$ and $\sigma_{\mathrm{I}} = (108 \pm 8)\,\mu\mathrm{m}$ respectively.

Derivation for a simplified model. To understand how the spatial kernel G_d enables one to solve equation 1 and to illustrate the key ideas behind our derivation of the linear response for the full model, we first consider a simplified model whose connectivity depends only on the cell type and spatial location of the pre- and post-synaptic neurons, and whose connectivity width depends only on pre-synaptic cell type. For this simplified model, the connectivity function is given by

$$W_{\alpha\beta}(\boldsymbol{x}-\boldsymbol{y}) = \frac{w_{\alpha\beta}}{\sigma_\beta^2} G_d(r;\sigma_\beta^{-2}). \qquad [3]$$

where $r = \|\boldsymbol{x}-\boldsymbol{y}\|$, and the division by σ_β^2 ensures that the integral of $W_{\alpha\beta}$ over space is $w_{\alpha\beta}$.

To solve for the system's linear response to a perturbation, we use the standard bra-ket notation (Materials and Methods) to rewrite equation 1 in a more abstract form

$$|r\rangle = W|r\rangle + |h\rangle \qquad [4]$$

where $|r\rangle, |h\rangle$ are the firing rate function and the perturbing input function respectively, and W is the linear integral operator that acts on $|r\rangle$ according to equation 1. The perturbation response vector can be written as $|r\rangle = (I-W)^{-1}|h\rangle$, so our goal is to compute the operator $L := (I-W)^{-1}$.

The most common approach, is to compute the perturbative expansion of the linear response operator in the form of a Neumann series $(I-W)^{-1} = \sum_{n=0}^{\infty} W^n$ (2, 21–25). However, this approach suffers from two key issues: 1) the series does not converge for operators W whose spectral radius is greater than 1 (Figure 1C), and 2) even when the series converges, the number of terms required for a good approximation may be large, thus failing to provide a simple description of the relationship between connectivity and perturbation response.

The choice of spatial kernel G_d, allows for exact computation of the inverse $L = (I-W)^{-1}$. This is because the definition of $G_d(\|\cdot\|;\sigma^{-2})$ as the Green's function of $\sigma^{-2} - \nabla^2$ allows us to write the connectivity operator W as

$$W = \boldsymbol{W}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1} - \nabla^2)^{-1} \qquad [5]$$

where $\boldsymbol{W}$ is the matrix of elements $w_{\alpha\beta}$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with elements σ_β^2. But by the Woodbury matrix (operator) identity (26), $(I - UC^{-1})^{-1} = I + U(C-U)^{-1}$ for any operators U, C. Thus, if we take $U = \boldsymbol{W}\boldsymbol{\Sigma}^{-1}$ and $C = \boldsymbol{\Sigma}^{-1} - \nabla^2$, and assume that $(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{\Sigma}^{-1}$ is diagonalizable as $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$, then

$$L = I + \boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{P}(\boldsymbol{\Lambda} - \nabla^2)^{-1}\boldsymbol{P}^{-1} \qquad [6]$$

As $\tilde{L} := L - I$ is analogous to the connectivity operator defined by equation 5, if we let $\tilde{L}_{\alpha\beta}(\boldsymbol{x}-\boldsymbol{y})$ be the response of neuron $(\alpha, \boldsymbol{x})$ to perturbation of a different neuron $(\beta, \boldsymbol{y})$, then $\tilde{L}$ can be written as

$$\tilde{L}_{\alpha\beta}(\boldsymbol{x}-\boldsymbol{y}) = \sum_{\gamma=0}^{N_c-1} [\boldsymbol{W}\boldsymbol{\Sigma}^{-1}\boldsymbol{P}]_{\alpha\gamma}[\boldsymbol{P}^{-1}]_{\gamma\beta} G_d(r;\lambda_\gamma) \qquad [7]$$

where λ_γ are the diagonal entries of $\boldsymbol{\Lambda}$, *i.e.* the eigenvalues of $(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{\Sigma}^{-1}$.

The full model. We define the connectivity function of the full space- and feature-dependent model by

$$\begin{aligned} &W_{\alpha\beta}(\mu,\nu,\boldsymbol{x}-\boldsymbol{y},\theta-\phi) \\ &= \frac{w_{\alpha\beta}}{2\pi\sigma_{\alpha\beta}^2} G_d(r;\sigma_{\alpha\beta}^{-2})(1 + 2\kappa_{\alpha\beta} f_\alpha(\mu) g_\beta(\nu)\cos(\theta-\phi)) \end{aligned} \qquad [8]$$

where $\kappa_{\alpha\beta} \in [-0.5, 0.5]$, and $f_\alpha, g_\alpha \in L^2([0,1])$ are monotonically increasing functions such that $f_\alpha(0) = g_\alpha(0) = 0$, $f_\alpha(1) = g_\alpha(1) = 1$. The sign of $\kappa_{\alpha\beta}$ determines whether connectivity is correlated or anti-correlated with difference in feature preference, while f_α and g_α determine the strength of this correlation as a function of feature selectivity. Under this choice of W, the response $r_\alpha(\mu, \boldsymbol{x}, \theta)$ to a single-cell perturbation of a different neuron $(\beta, \nu, \boldsymbol{y}, \phi)$ can be found to be (SI section 2)

$$\begin{aligned} &\tilde{L}_{\alpha\beta}(\mu,\nu,\boldsymbol{x}-\boldsymbol{y},\theta-\phi) \\ &= \frac{1}{2\pi}\left(\tilde{L}_{0\alpha\beta}(r) + 2\tilde{L}_{1\alpha\beta}(r) f_\alpha(\mu) g_\beta(\nu)\cos(\theta-\phi)\right) \end{aligned} \qquad [9]$$

where the definition of $\tilde{L}_{n\alpha\beta}(r)$ (equation 13) has a similar form to equation 7, generalized to allow connectivity widths to depend on both pre- and post-synaptic cell types and to include feature preference dependence.

Since equation 9 is exact, we should expect a close agreement between our theory and numerical simulations of the model regardless of the spectral radius of the connectivity matrix. Indeed, we obtain near perfect agreement between our theory and numerical simulations for the single cell perturbation response in an E-I model with two spatial dimensions and a spectral radius of 1.8 (Figure 1D; Materials and Methods). For comparison, we also computed the perturbation response using the Neumann series expansion of the matrix inverse up to 3rd order (*i.e.* $L \approx \sum_{n=0}^{3} W^n$). This is the minimum order at which the responses of excitatory neurons to the perturbation of a single excitatory neuron depend on all connectivity weights (including I → I weights). As expected, the series expansion severely diverges from simulations due to the spectral radius being greater than 1 (Figure 1D).

Mean response of unperturbed neurons. Perturbation of a single pyramidal neuron results in mean suppression of unperturbed neurons (1), suggesting that inhibitory connections are sufficiently strong in order to overcome recurrent excitation. However, the precise conditions under which mean suppression occurs are unclear. To address this question, we integrate equation 9 over all its continuous variables to obtain an expression for the mean response of unperturbed neurons to single-cell perturbations, given by $\tilde{\boldsymbol{L}} = (\boldsymbol{I}-\boldsymbol{W})^{-1} - \boldsymbol{I}$, where $\tilde{L}_{\alpha\beta}$ is the mean response of cell type α to perturbation of cell type β (SI section 3). In the specific case of an E-I model, it can be shown that for single-cell excitatory neuron perturbations, unperturbed excitatory neurons are suppressed on average if and only if $\det(\boldsymbol{W}) > w_{\mathrm{EE}}$, or equivalently, $|w_{\mathrm{EI}}|w_{\mathrm{IE}} > w_{\mathrm{EE}}(|w_{\mathrm{II}}|+1)$, while inhibitory neurons are always excited on average (SI section 6). Thus the observation of mean suppression of unperturbed neurons implies that the disynaptic E → I → E inhibition must be stronger than the product of E → E excitation and I → I inhibition.

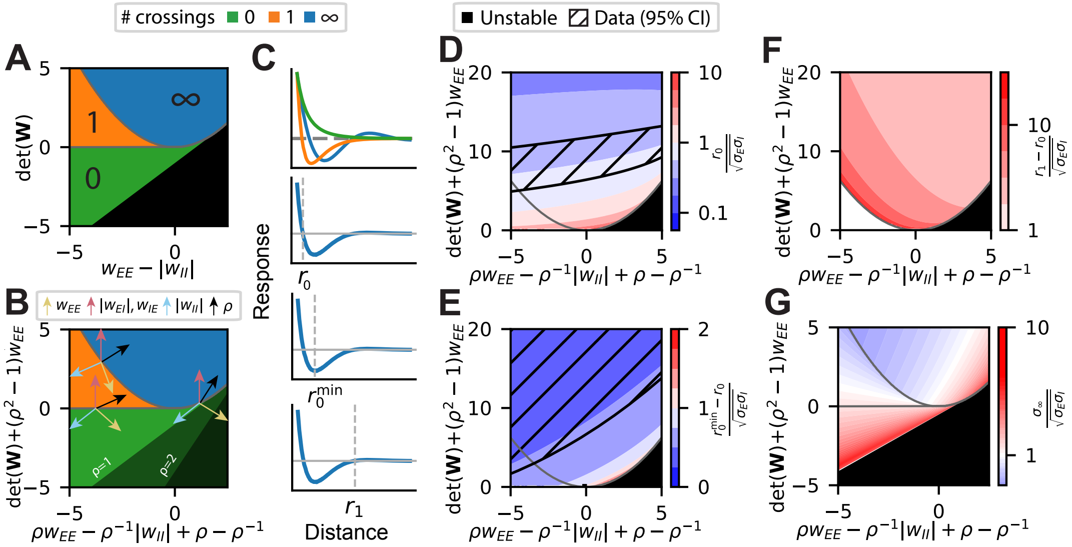

Fig. 2. Spatial profile of excitatory neuron response to single-cell perturbation in E-I ISNs. A) Phase diagram of the number of zero crossings in the perturbation response as a function of distance from the perturbation for networks with ρ=1 ($i.e.$ $\sigma_E = \sigma_I$). Networks in the phase region shaded in black are dynamically unstable. The phase boundaries between 0, 1, and ∞ are given by $y = 0$ and $y = \frac{x^2}{4}$. **B)** Phase diagram of the number of zero crossings for networks with arbitrary ρ. The instability region is dependent on ρ, with boundary $y = \rho(x - \rho)$ for $x \leq 2\rho$ and $y = \frac{x^2}{4}$ for $x > 2\rho$. Arrows indicate changes in number of zero crossings induced by perturbations of each parameter at the phase boundaries. **C)** Top panel: Illustration of the perturbation response as a function of distance within each of the three phase regions. Remaining panels: Illustration of the quantities r_0, $r_0^{\min}$, and r_1 as plotted in D-F. **D)** Location to the first zero-crossing, r_0, as a fraction of the connectivity length scale $\sqrt{\sigma_E \sigma_I}$ for 2-dimensional models with $w_{\mathrm{EE}} = 5$, $\rho = 0.72$. 95% confidence interval of $\frac{r_0}{\sqrt{\sigma_E \sigma_I}}$ estimated from experimental data (1, 11, 12; Materials and Methods) is indicated by hatched region. Grey line indicates the boundary between 1 and ∞ zero crossings as seen in A and B. **E)** Similar to D, but the distance from the first zero crossing r_0 to the first minimum $r_0^{\min}$ is plotted. **F)** Similar to D, but the distance from the first zero-crossing r_0 to the second zero-crossing r_1 is plotted. **G)** Asymptotic decay length scale σ_∞ for models with $\rho = 0.72$. Note that unlike D-F, this variable is independent of the specific choice of w_{EE} and the number of spatial dimensions d. Panels D-F are computed for 2-dimensional models; panels A and B are valid for 2 or more dimensions.

Spatial profile of perturbation response. In addition to the mean suppression of unperturbed neurons, single-cell perturbations of pyramidal neurons produce a Mexican-hat-shaped response as a function of distance, where neurons near the perturbed site are excited and neurons farther away are suppressed (1). Intuitively, this would suggest a connectivity motif of narrow excitation and broad inhibition. However, recent mouse V1 L2/3 connectivity data shows that the opposite is true: E → I and I → E connections are narrower than E → E connections (11, 20). Furthermore, the length scale of E → E connectivity (standard deviation ≈ 125 µm for a Gaussian spatial profile, 20) is significantly broader than the spatial radius of nearby excitation (≈ 70 µm, 1), and an even shorter radius of excitation (≈ 35 µm) is seen for multi-cell perturbations, which could not be explained by a model with a Gaussian spatial profile for each connection (2). Thus we set out to investigate the conditions under which such small radii of nearby excitation can arise in our model with realistic connectivity length scales.

Number of spatial zero crossings. The Mexican-hat-shaped spatial profile of perturbation response implies that the response crosses zero from nearby activation to suppression at least once, or in other words, that there is at least one zero crossing in the response as a function of distance from the perturbation. It is conceivable that the response changes sign more than once, but that these zero crossings cannot be detected due to measurement noise. Thus, the question of whether or not the model can exhibit the Mexican-hat-shaped profile of perturbation response can be broken into two mathematical sub-problems: whether or not nearby neurons are activated, and whether or not there exists at least one zero crossing in the response as a function of distance. We find that for all networks with 2 or more spatial dimensions, single-cell excitatory neuron perturbations always activates nearby neurons, in the mathematical sense that neurons arbitrarily close to the perturbed cell are activated (SI section 7).

To proceed further, we assume that the connectivity width depends only on pre-synaptic cell type. In this case, we find that E-I models may exhibit either 0, 1, or infinitely many zero crossings (SI section 8A). The exact behavior is determined by both the connectivity width and connectivity strength via the two eigenvalues λ_γ of the 2×2 matrix $(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{\Sigma}^{-1}$. If λ_0, λ_1 are complex conjugates, then the response of both excitatory and inhibitory neurons must exhibit infinitely many zero crossings. If λ_γ are real and the network is an ISN with two or more spatial dimensions, the condition for excitatory neuron response having exactly one zero crossing is that the smaller of the two eigenvalues, λ_0, satisfy $\lambda_0 > \sigma_{\mathrm{E}}^{-2}$, and the same condition for inhibitory neurons is $\lambda_0 > \sigma_{\mathrm{I}}^{-2}$ (SI Corollary 8.4). Thus, not only can the model exhibit the Mexican-hat-shaped profile of perturbation response, but we are also able to determine the precise conditions under which this occurs.

The mathematical conditions on the number of zero crossings can be formulated more intuitively in terms of the connectivity strengths $w_{\alpha\beta}$ and the ratio of inhibitory to excitatory connectivity width $\rho = \frac{\sigma_{\mathrm{I}}}{\sigma_{\mathrm{E}}}$. Note that those conditions, and the following results presented in Figures 2 and 3, assume that the E-I network is an ISN with two or more spatial dimensions whose connectivity widths depend only on presynaptic cell type. We first consider the special case in which the inhibitory and excitatory spatial kernels have the same width ($\rho = 1$). In this case, the number of zero crossings of excitatory neuron response can be represented as a phase diagram in terms of the trace and determinant of $\boldsymbol{W}$ (Figure 2A). This diagram reveals some simple principles governing the number of zero crossings: First, the existence of at least one zero-crossing implies $\det(\boldsymbol{W}) = |w_{\mathrm{EI}}|w_{\mathrm{IE}} - w_{\mathrm{EE}}|w_{\mathrm{II}}|$ must be positive, and hence the disynaptic E → I → E inhibition must be stronger than product of E → E and I → I connections. Second, notice that I → I connections must be stronger than E → E (so that the value on the x-axis is less than 0) for the network response to exhibit exactly one zero crossing. This suggests I → I connections may have the regularizing role of suppressing spatial oscillations.

Our phase diagram can be generalized to the case of arbitrary ρ by modifying the axes (Figure 2B; SI section 8B). To gain intuition about this phase diagram, we analyze the change in the number of zero crossings induced by increasing each of the connectivity parameters at the phase boundaries (SI section 8C), as indicated by the colored arrows in the figure. We find that increasing ρ (*i.e.* broadening inhibitory connections) encourages the formation of zero crossings, as one would expect intuitively. Increasing the strength of E → I and I → E connections also encourages the formation of zero crossings, while increasing the strength of E → E and I → I connections has the opposite effect. Furthermore, the phase diagram reveals that the principles we obtained for the case of $\rho = 1$ can be generalized with slight modifications: First, the existence of at least one zero crossing implies the determinant $\det(\boldsymbol{W})$ must be greater than $(1 - \rho^2)w_{\mathrm{EE}}$, which is positive for networks with $\rho < 1$. Second, for the network response to exhibit exactly one zero crossing, I → I connections must be stronger than $\rho^2 w_{\mathrm{EE}} + (\rho^2 - 1)$, which in turn must be stronger than E → E connections if $\rho > 1$.

Spatial radius of nearby excitation. We have shown that our model can qualitatively exhibit the Mexican-hat response to excitatory perturbations found in data (1), given sufficiently strong disynaptic E → I → E inhibition. However, the location of the first zero crossing (*i.e.* the spatial radius of nearby excitation), r_0, has been measured at approximately 70 µm (1), which is significantly narrower than the connection probability length scale at around 100 – 125 µm (11, 20). Can this be explained by our model? To address this we compute r_0 at different points of the phase space as a fraction of the geometric mean of the connectivity length scales, $\sqrt{\sigma_E \sigma_I}$. Since this quantity is not fully determined by the x- and y-axes of Figure 2B, we compute it for different combinations of w_{EE} and ρ (Figure S1). The specific case of $w_{\mathrm{EE}} = 5, \rho = 0.72$ is illustrated in Figure 2D, where the value of 0.72 is our best estimate of ρ obtained from the fitted connectivity kernels in Figure 1B. These numerical results show that r_0 is negatively correlated with $\det(\boldsymbol{W})$, such that the determinant must be considerably greater than 0 (*i.e.*, disynaptic E → I → E inhibition must be significantly stronger than the product of E → E and I → I connections) in order to explain the narrow Mexican-hat-shaped response profile observed by (1). Note that this condition is more stringent than the condition $\det(\boldsymbol{W}) > 0$ for the existence of at least one zero crossing.

Spatial location of maximum suppression.. Further constraints on the connectivity parameters can be inferred by considering the distance to the first local minimum $r_0^{\min}$ of the perturbation response, which we expect to be the spatial location of maximum suppression. Unlike the location of the first zero crossing r_0, the additional distance to the first minimum, $r_0^{\min} - r_0$, is moderately invariant to the specific choice of w_{EE} and ρ (Figure S2; Figure 2E shows the specific case of $w_{\mathrm{EE}} = 5, \rho = 0.72$). Combined with the observation that the contour lines of $r_0^{\min} - r_0$ are diagonal, this implies a correlation between the values of $\det(\boldsymbol{W})$ and $\mathrm{tr}(\boldsymbol{W})$ that can explain the data.

Experimental data places $r_0^{\min}$ at around 110 µm (1), so that $r_0^{\min} - r_0$ is around 40 µm, which is less than half of the connectivity width length scale of $\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}} \approx 127$ µm as measured from Figure 1B. This would place the network in the darker blue region – roughly, the upper left triangle – of Figure 2E, which overlaps considerably with the appropriate region of Figure 2D as determined above.

Frequency of spatial oscillations. As we have shown, the region of phase space with only one zero-crossing requires sufficiently strong I → I inhibition (Figure 2B). This suggests that I → I inhibition is important for suppressing spatial oscillations. This intuition can be made precise by considering the distance from the first zero-crossing r_0 to the second zero-crossing r_1, a quantity that is invariant to the choice of w_{EE} in one and three spatial dimensions (SI section 9A), and almost invariant in two dimensions (Figure S3). As expected from the intuition, $r_1 - r_0$ increases (*i.e.* frequency of spatial oscillations decreases) with the strength of I → I inhibition (Figure 2F). More precisely, it can be proven that in one or three spatial dimensions, the derivative of $r_1 - r_0$ with respect to $|w_{\mathrm{II}}|$ is always positive (SI section 9B).

Stability and spatial decay length scale. Finally we consider the rate at which the perturbation response decays with distance. Since the response is a non-monotonic function of distance, we measure its asymptotic decay length scale σ_∞, defined such that the perturbation response decays asymptotically as $r^{-\frac{d-1}{2}} e^{-\frac{r}{\sigma_\infty}}$ as $r \to \infty$. Under the assumption of fast inhibition, we find an interesting relationship between σ_∞ and the overall stability of the network: the closer the network is to the edge of instability, the longer the decay length scale (Figure 2G, SI section 11). This relationship is fully general, applying to networks with arbitrary number of cell types and arbitrary connectivity widths and spatial dimensions. Thus, assuming sufficiently fast inhibition, observation of a decay length scale of the same order of magnitude as, or smaller than, the connectivity length scale would suggest that the network is reasonably far from the edge of instability.

Inhibitory neuron response. Thus far we have focused on the responses of excitatory neurons to perturbations. This is because, to the best of our knowledge, existing simultaneous two-photon optogenetics and calcium imaging experiments in mouse V1 either do not discriminate between the responses of excitatory and inhibitory neurons, or only measure the responses of excitatory neurons (1–3, 5). However, the responses of inhibitory neurons encode important information about the recurrent connectivity: for example, whether the cortical circuit is an ISN can be determined by a paradoxical effect whereby inhibitory neurons are suppressed by optogenetic stimulation of inhibitory neurons (8–10, 27, 28).

We find that, in response to perturbation of a single excitatory cell, the responses of inhibitory neurons are tightly related to those of excitatory neurons. Consider, again,

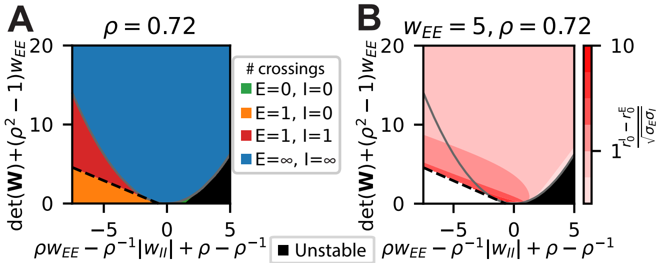

Fig. 3. Relationship between the spatial profile of excitatory and inhibitory neuron response to single-cell perturbation in E-I ISNs with 2 or more spatial dimensions. A) Phase diagram of the number of E and I zero crossings for networks with $\rho = 0.72$. Green: Neither E nor I exhibit zero crossings. Orange: E exhibits one zero crossing, I exhibits no zero crossing. Red: Both E and I exhibit one zero crossing. Blue: Both E and I exhibit infinitely many zero crossings. Dashed line is given by the equation $y = (\rho - \rho^{-1})(x - (\rho - \rho^{-1}))$. **B)** Distance between the first zero crossing of excitatory neuron response and the first zero crossing of inhibitory neuron response in networks with two spatial dimensions and $w_{\mathrm{EE}} = 5, \rho = 0.72$.

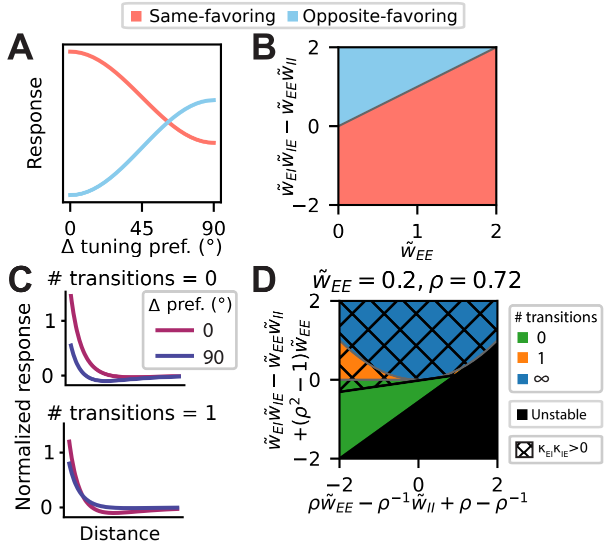

Fig. 4. Feature-tuning dependence of excitatory neuron response to single-cell perturbation in E-I networks. A) Illustration of same-favoring and opposite-favoring responses. **B)** Phase diagram of feature tuning of perturbation response, with red indicating same-favoring response (iso-tuned neurons are more excited than ortho-tuned neurons), and blue indicating opposite-favoring response (the opposite of same-favoring).**C)** Example responses of networks with 0 and 1 transitions between same- and opposite-favoring response with increasing distance, normalized for visual clarity. **D)** Phase diagram of number of such transitions for a two-dimensional network with $\tilde{w}_{\mathrm{EE}} = 0.2$ and $\rho = 0.72$. Green, orange, and blue represent 0, 1, and ∞ transitions respectively, while black represent region of instability. Hatched region indicates like-to-like disynaptic E → I → E inhibition. Orange and blue regions are contained within the hatched region, showing that the presence of at least one transition implies like-to-like disynaptic inhibition.

an E-I ISN in two or higher dimensions with connectivity widths depending only on pre-synaptic cell type. We can show that 1) the excitatory neuron response is oscillatory (having an infinite number of zero crossings as a function of distance) if and only if inhibitory neuron response is also oscillatory, and 2) if excitatory neuron response exhibits a single zero-crossing as a function of distance, then inhibitory neuron response must also exhibit a single zero-crossing unless $\sigma_{\mathrm{E}}^{-2} < \lambda_0 < \sigma_{\mathrm{I}}^{-2}$ (SI section 8E). These relations are summarized by the phase diagram in Figure 3A. Thus, for most parameter regimes we expect the responses of inhibitory neurons to exhibit the same number of zero-crossings as excitatory neurons. Violations of this expectation, however, would suggest that the connection strengths between E and I satisfy tight inequalities.

Now suppose that inhibitory neurons indeed exhibit a Mexican-hat shaped response profile. As explained in the section Mean response of unperturbed neurons, mean inhibitory neuron response must be positive. Given the mean suppression of excitatory neurons, we thus expect less lateral suppression of inhibitory neurons than excitatory neurons. In particular, it can be shown that if, and only if, $\det(\boldsymbol{W}) > 0$, the inhibitory response profile has a greater spatial radius of nearby excitation than the excitatory response profile, that is, $r_0^{\mathrm{I}} > r_0^{\mathrm{E}}$, where $r_0^{\mathrm{E}}, r_0^{\mathrm{I}}$ are the distances to the first zero crossing of excitatory and inhibitory neuron responses respectively (SI section 9D). This is illustrated by Figure 3B for the case of $w_{\mathrm{EE}} = 5, \rho = 0.72$. Other combinations of w_{EE} and ρ are shown in Figure S4. Note that $r_0^{\mathrm{I}} > r_0^{\mathrm{E}}$ for all subplots with $\rho \leq 1$ since for these networks, existence of a zero crossing implies $\det(\boldsymbol{W}) > 0$ (Figure 2B). Furthermore, recall that there is mean suppression of excitatory neurons if and only if $\det(\boldsymbol{W}) > w_{\mathrm{EE}}$. Thus, given mean suppression of excitatory neurons, inhibitory neuron response must be less suppressed and exhibit a broader spatial profile than excitatory neuron response.

Feature-tuning dependence of perturbation response. Upon optogenetic perturbation of a single excitatory neuron, neurons in L2/3 of mouse V1 that have tuning similar to that of the perturbed neuron (iso-tuned neurons) are, on average over space, more suppressed than neurons that have orthogonal tuning (ortho-tuned neurons) (1). We call this an *opposite-favoring* response, as opposed to a *same-favoring* response in which iso-tuned neurons are less suppressed or more excited than ortho-tuned neurons (Figure 4A). Since E → E connectivity in L2/3 of mouse V1 is *like-to-like*, meaning similarly tuned excitatory neurons are preferentially connected (11, 29), this suggests the need for a like-to-like disynaptic E → I → E inhibition motif to obtain preferential suppression of similarly tuned excitatory neurons.

To determine if this intuition is correct, we integrate equation 9 over space to obtain the average perturbation response as a function of feature tuning (SI equation S32). Given like-to-like E → E connectivity, we find that excitatory neuron response is opposite-favoring if and only if $\tilde{w}_{\mathrm{EI}}\tilde{w}_{\mathrm{IE}} > \tilde{w}_{\mathrm{EE}}(\tilde{w}_{\mathrm{II}} + 1)$ (Figure 4B; SI section 12A), where $\tilde{w}_{\alpha\beta} = |w_{\alpha\beta}|\kappa_{\alpha\beta}\int_0^1 f_\alpha(\mu)g_\alpha(\mu)P_\alpha(\mu)\,d\mu$ is positive if and only if the connectivity from cell type β to α is like-to-like, *i.e.* $\kappa_{\alpha\beta}$ is positive. Under this condition, like-to-like E → I → E inhibition ($\kappa_{\mathrm{EI}}\kappa_{\mathrm{IE}} > 0$) is not necessary if $\tilde{w}_{\mathrm{II}} < -1$. However, networks with $\tilde{w}_{\mathrm{II}} < -1$ and anti-like-to-like E → I → E inhibition ($\kappa_{\mathrm{EI}}\kappa_{\mathrm{IE}} \leq 0$) are unstable (SI section 12B). Thus, the observation of opposite-favoring response implies that disynaptic E → I → E connections provide like-to-like inhibition.

Modulation of feature-tuning dependence by distance. The single-cell perturbation response measured experimentally is not only opposite-favoring on average, it is opposite-favoring at all distances beyond 25 µm, if one computes tuning similarity as signal correlation (1). We find that in models with two or more spatial dimensions and like-to-like E → E connections, sufficiently nearby excitatory neurons always exhibit same-favoring response (SI section 13). Thus, in order to explain the data, our model should exhibit a very nearby transition from same-favoring to opposite-favoring response with increasing distance from the perturbed neuron (Figure 4C). Indeed, E-I networks with two or more spatial dimensions and connectivity width that depends only on presynaptic cell type can exhibit 0, 1, or ∞ number of transitions between same- and opposite-favoring response (SI section 13). The number of such transitions is determined by combinations of $\tilde{w}_{\alpha\beta}$ and ρ (Figure 4D). Interestingly, given like-to-like E $\to$ E connectivity and the presence of at least one such transition (which is required to explain the data), disynaptic E $\to$ I $\to$ E connectivity must be like-to-like (Figure 4D; SI Theorem 13.4). In other words, given that sufficiently nearby neurons have same-favoring responses, if the perturbation response is opposite-favoring at any distance, then the disynaptic E $\to$ I $\to$ E inhibition must be like-to-like. Note that this finding is stronger than our previous finding that a response whose mean across distance is opposite-favoring implies like-to-like E $\to$ I $\to$ E connectivity.

Modulation of perturbation response by neuronal gain. While perturbation of a single pyramidal neuron leads to an opposite-favoring response (1), perturbation of an ensemble of 10 similarly-tuned pyramidal neurons results in a same-favoring, rather than opposite-favoring, response (2). There are three important differences between these experiments that could underlie these seemingly contradictory results. One difference is the number of stimulated cells. Second, the single cell perturbation experiment measured all cells, while the ensemble perturbation experiment measured only E cells. A scenario in which excitatory neurons exhibit weakly same-favoring response and inhibitory neurons exhibit strongly opposite-favoring response, such that the average of E and I response is opposite-favoring, could therefore explain both results. However this seems unlikely since most neurons in the cortex are excitatory. The third difference, which we address here, is that the two experiments were performed under different stimulus conditions: the single-cell perturbation was performed with the simultaneous presentation of a visual stimulus (drifting gratings), and thus with a higher background firing rate, while the ensemble perturbation experiment was performed with only a gray screen. If cortical cells have supralinear input/output functions (30–32, but see 33), then their gain – the change in rate for a given change in input – would be increased for higher firing rates. This in turn would increase the effective connection strengths which, in a model linearized about a fixed point, are given by the gains times the synaptic weights. This increased gain and increased connectivity strength might explain the difference between the two experiments. Motivated by this reasoning, we study how various perturbation response properties are modulated by neuronal gain.

Modulation of mean perturbation response by neuronal gain. First we study how changes in neuronal gain (g), which in our model effectively scales all connectivity weights by g, modulate the mean response. We find that if the unperturbed excitatory neurons exhibit mean suppression, then increasing neuronal gain always results in stronger suppression (SI section 6). Similarly, reducing neuronal gain always results in weaker suppression or, for sufficiently small gain, mean excitation (Figure 5A). Note that the derivative of the mean response with respect to gain is non-monotonic, such that if the unperturbed excitatory neurons exhibit mean excitation, then increasing the gain may result in stronger excitation instead.

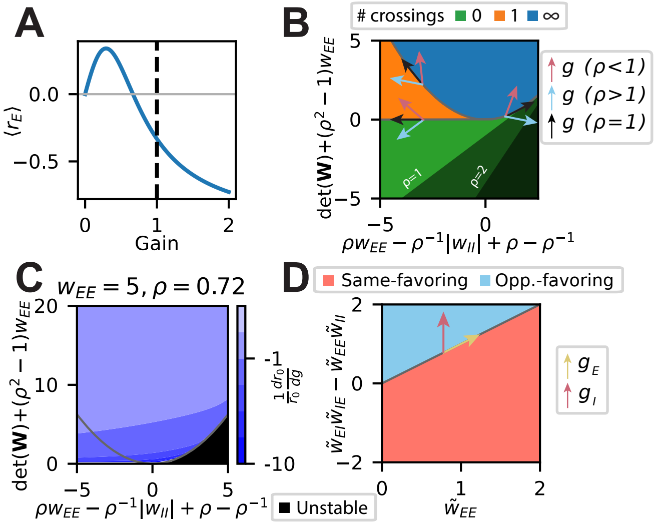

Fig. 5. Modulation of excitatory neuron response to single-cell perturbation by neuronal gain in E-I ISNs with 2 spatial dimensions A) Mean response of unperturbed excitatory neurons as a function of gain, for a network with $w_{\mathrm{EE}} = 2, w_{\mathrm{II}} = -1, \det(\boldsymbol{W}) = 5$. **B)** Phase diagram of number of zero crossings from Figure 2B. Arrows indicate changes in number of zero crossings induced by increasing gain at the phase boundaries for $\rho = 1$, $\rho > 1$, or $\rho < 1$. **C)** Derivative of distance to the first zero crossing with respect to gain, divided by the distance, for $w_{\mathrm{EE}} = 5, \rho = 0.72$. **D)** Phase diagram of feature tuning of perturbation response, with red indicating same-favoring response and blue indicating opposite-favoring response. Arrows indicate movement in phase space induced by increasing excitatory and inhibitory neuron gain respectively at the phase boundary.

Modulation of the spatial profile of the response by neuronal gain. Next, we study the modulation of the number of spatial zero crossings by neuronal gain. We find that the changes in the number of spatial zero crossing due to increasing gain depend entirely on the value of ρ (Figure 5B; SI section 8D): if $\rho = 1$, then an increase in gain does not change the number of spatial zero crossings of the response; while if $\rho < 1$ or $\rho > 1$, then, if starting from near a phase boundary, increasing gain increases or decreases, respectively, the number of zero crossings.

We next study the effect of gain on the location of zero crossings. We compute the derivative of the distance to the first zero crossing r_0 with respect to the gain g, and find that when $\rho = 1$, the derivative is always negative (SI section 9C). This means that if $\rho = 1$ and a zero crossing exists, then increasing the gain always produces a narrower spatial radius of nearby excitation. Numerically, we find that this also holds when $\rho < 1$ (Figure 5C), and is mostly true when $\rho > 1$ (Figure S5). Thus, given our estimate of $\rho \approx 0.72$ in experimental data, we predict that single-cell perturbation experiments performed while presenting only a grey screen, which have a lower gain, should result in a broader response profile with less suppression and the same or a decrease in number of zero crossings.

Modulation of feature dependence by neuronal gain. We return to our original motivation for studying the gain modulation of perturbation responses: can a difference in gain explain the seemingly contradictory results reported regarding the feature dependence of perturbation response? We find that increasing gain may result in a transition from same-favoring to opposite-favoring excitatory neuron response (SI section 12C). Furthermore, if we selectively increase the gain of excitatory or inhibitory neurons only, we find that this transition is mediated by an increase in gain of inhibitory neurons g_I (*i.e.* effective scaling of all connections onto inhibitory neurons by g_I), whereas increasing the gain of excitatory neurons g_E cannot yield such a transition (Figure 5D). More importantly, it can be shown that a transition from opposite- to same-favoring response can always be induced by sufficiently decreasing the gain of inhibitory neurons (SI section 12C). Thus, the difference in neuronal gain may indeed be the explanation for why an opposite-favoring response is observed in the experiment with drifting grating stimuli (1) while a same-favoring response is observed in the experiment without visual stimuli (2), suggesting that a supralinear transfer function of neurons (or at least of inhibitory neurons) may be important for switching between two qualitatively distinct computations.

Validation of theoretical insights in fitted models

So far all of our theoretical analysis of the properties of the linear response function has relied on two simplifying assumptions. First, we have assumed, for theoretical tractability, that the connectivity width depends only on the presynaptic cell type, *i.e.* the models obey the symmetries $\sigma_\mathrm{EE} = \sigma_\mathrm{IE}$ and $\sigma_\mathrm{EI} = \sigma_\mathrm{II}$. However, recent connection probability data from L2/3 of mouse V1 suggests that this symmetry may not hold, and that the connection probability length scales instead satisfy the relations $\sigma_\mathrm{EE} \approx \sigma_\mathrm{II} > \sigma_\mathrm{EI} \approx \sigma_\mathrm{IE}$ (20). Second, we have analyzed the single-cell perturbation response in (1) under the simplifying assumption that the measured responses are all of excitatory neurons, while in the experiment both excitatory and inhibitory neurons were measured. To test the robustness of our findings, we relax these assumptions and fit models so that the mix of 85% excitatory and 15% inhibitory cells match the perturbation response from (1), both as a function of distance and as a function of orientation tuning preference. We also constrain the models to have the parameter κ_{EE} within two standard deviations of our estimate from data of (11) (Materials and Methods).

From the 200 fitted models, we select the 50 best-fitting models for analysis (Materials and Methods). Consistent with our theoretical analysis of the spatial profile of perturbation response, all fitted models exhibit a positive determinant of the weight matrix $\boldsymbol{W}$, and the determinant and trace of $\boldsymbol{W}$ are correlated across models (Figure 6A; compare Figure 2D, E). Most fitted models (47/50) also exhibit like-to-like disynaptic E → I → E inhibition as suggested by our theory (Figure 6B; compare Figure 4D). Furthermore, we find that the two exceptions nonetheless confirm our prediction that negative $\kappa_{IE}\kappa_{EI}$ implies same-favoring excitatory responses; these two cases follow the unlikely scenario we referred to previously, in which the same-favoring behavior of the excitatory cells is weak enough, and the opposite-favoring behavior of the inhibitory cells strong enough, that the average over the population matches the opposite-favoring behavior of the data. Despite large variances in model parameters, the perturbation responses of all the fitted models closely match experimental data (Figure 6C, D). On average, the fitted models display opposite-favoring responses across almost the entire range of experimentally measured distances (Figure 6E), consistent with the findings of (1).

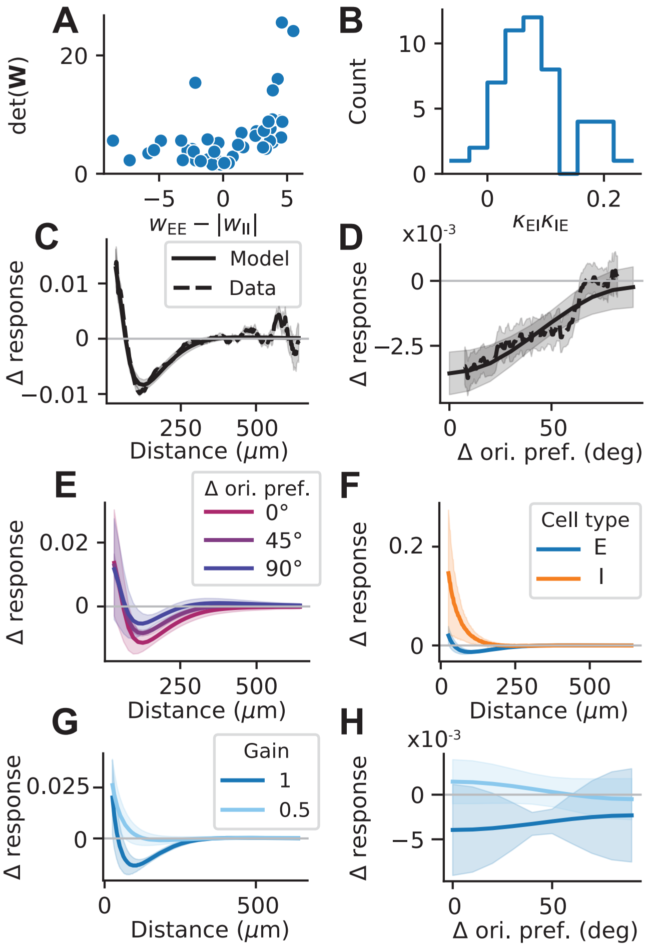

Fig. 6. Validation of theoretical insights in fitted models. 200 models are fitted to the single cell-perturbation response curve as a function of distance from (1), the top 50 of which are plotted. **A)** Distribution of the fitted model parameters, where each point is a fitted model. **B)** Histogram of the product of fitted parameters $\kappa_\mathrm{EI}\kappa_\mathrm{IE}$, which is positive if and only if disynaptic E → I → E inhibition is like-to-like. **C-E)** Perturbation response of all neurons in the model (including E and I) **C-D)** Comparison between the perturbation response of the fitted models and experimental data. Error bars of data represents standard error. Error bars of model represents standard deviation across fitted models. To match the data analysis procedure of (1), bin widths of $60\,\mu\mathrm{m}$ for C and 25° for D are used. **E)** Perturbation response of fitted models as a function of distance to the perturbed neuron, for different tuning preferences. Same bin width as C). **F-H)** Simulations support analytical predictions. Smaller bin widths than C-E are used for more accurate results ($2\,\mu\mathrm{m}$ bins for F, G and 10° bins for H). **F)** Comparison between excitatory and inhibitory neuron response in fitted models. **G-H)** Effect of reducing neuronal gain on the responses of excitatory neurons. Models are fitted with a gain of 1.

We then test three of our theoretical predictions on these fitted models: 1) inhibitory neurons should exhibit a broader perturbation response profile than excitatory neurons (Figure 6F), 2) when overall neuronal gain is lowered, excitatory neuron response should be broader and less suppressed (Figure 6G), and 3) when overall neuronal gain is sufficiently weak, excitatory neuron response transitions from opposite-favoring to same-favoring (Figure 6H). These predictions hold true in all the fitted models, despite the large variances in model parameters and despite the fact that these models violate the symmetry assumptions in our theory, suggesting that these are robust effects that can be expected from experimental measurements.

Discussion

In this paper, we developed novel theory for understanding the link between recurrent connectivity structure and single-cell optogenetic perturbation responses. We introduced an exponential-type kernel for describing connectivity as a function of distance that for the first time allows an exact solution for a space- and feature-dependent linear network that is valid in all coupling regimes. We showed that this kernel can well capture the spatial dependence of the connectivity in the data, defined as the product of the connection probability and its strength, and used this to exactly solve for the network's steady-state response to a single-cell perturbation.

Analysis of the solution for the class of E-I ISN networks revealed five main results. First, we found that a positive determinant of the 2 × 2 connectivity weight matrix $\boldsymbol{W}$ is necessary (assuming inhibitory projections are narrower than excitatory) to explain experimental observations of a perturbation response that is excitatory for nearby cells and suppressive at larger distances. The larger the determinant, the shorter the spatial radius of nearby excitation. Second, we found that the response at larger distances can either remain negative or be oscillatory in space, and spatial oscillation frequency is negatively correlated with the strength of I → I connections. Third, we predicted that the spatial profile of the perturbation responses of inhibitory neurons qualitatively matches that of excitatory neurons, but that inhibitory neurons exhibit a larger spatial radius of nearby excitation than excitatory neurons. Fourth, examining dependence on feature tuning, we found that feature-specific disynaptic inhibition (E → I → E) that is like-to-like (*i.e.*, that couples neurons with similar preferred features) is necessary to explain experimental observations. These observations show that neurons with feature preferences opposite to a perturbed neuron are less suppressed or more excited on average than neurons with similar feature preferences, a phenomenon we called "opposite-favoring responses". In fact, such a like-to-like connectivity motif is necessary if the perturbation response is opposite-favoring at any distance. Finally, we predicted that a decrease in neuronal gain would cause perturbation response to be less suppressive and have a broader spatial radius of excitation, and that the response becomes same-favoring rather than opposite-favoring for sufficiently weak neuronal gain. All of the analytic results listed above except the fourth were obtained on the assumption that connectivity width depends only on presynaptic cell type. However, we found that our theoretical predictions hold in simulations without this assumption.

To the best of our knowledge, this is the first exactly solvable model of a recurrent network with space- and feature-dependent recurrent connectivity. Consider models that are "translation-invariant", meaning that connectivity depends only on spatial distance and difference in preferred feature, as well as on cell type (the model we study also includes a non-translation-invariant dependence on feature selectivity). It is straightforward to obtain an exact analytic solution of a linear translation-invariant model in Fourier space, but in general this cannot be inverted to obtain responses as a function of distance. Nonetheless, previous works were able to obtain some information analytically, *e.g.* using the Fourier-space solutions to compute the spatial resonant frequencies of the network, from which experimental predictions were made (13). Alternatively, one may obtain an approximate expression for the steady-state solution by assuming that all activity patterns have a Gaussian shape (18, 34), although this assumption, typically applied to visual responses, may not be suitable for describing single-cell perturbation responses. Ref. (14), obtained an exact steady-state solution for an E-I network with a Gaussian spatial connectivity kernel in the tightly balanced regime (14). In this regime, there is a precise cancellation between excitatory and inhibitory synaptic input currents such that $W|r\rangle + |h\rangle \approx 0$, so the steady state solution can be approximated as $|r\rangle \approx -W^{-1}|h\rangle$. However, experimental evidence suggests that the cortex is in a loosely balanced rather than a tightly balanced regime (35), and our model is valid in both regimes.

The exponential-type kernel we introduced for modeling the spatial dependence of connectivity is a natural higher-dimensional generalization of the exponential kernel for a 1D ring network studied by (36). Compared to the Gaussian kernel typically used for modeling mouse V1 connectivity (13–19), it has a sharp peak at short distances. This property of our spatial kernel satisfies the conditions recently found necessary to explain the short spatial radius of nearby excitation in perturbation responses (2), namely that this cannot be explained by models with spatial connectivity given by a single Gaussian kernel with realistic length scales, and that a sharp peak must be added to the connectivity kernel to explain the data.

A surprising corollary of our analysis of the spatial profile of perturbation responses is that, given an exponential-type connectivity kernel, a narrow perturbation response does not necessitate a narrow spatial connectivity kernel, and, conversely, neither does a narrow spatial connectivity kernel imply a narrow perturbation response. Instead, the spatial profile of perturbation response is strongly dependent on the mean connectivity strengths between different cell types. For example, Figure S1 shows that, given fixed connectivity widths, the spatial radius of nearby excitation can vary over several orders of magnitude depending on the E → E connectivity strength and the determinant $\det(\boldsymbol{W})$.

Our analysis of mean perturbation response as well as the spatial profile of perturbation response both strongly suggest the determinant $\det(\boldsymbol{W})$ is positive, *i.e.* the disynaptic E → I → E inhibition is stronger than the product of E → E and I → I connections. This has important implications for the network dynamics in a nonlinear E-I network. Because the linearized dynamics of a nonlinear network around the fixed point are driven by an effective connectivity matrix $\boldsymbol{W}$ equal to the product of the connectivity $\boldsymbol{J}$ and a diagonal matrix of (positive) neuronal gains, the determinant of $\boldsymbol{W}$ and $\boldsymbol{J}$ have the same sign. Thus, our insight that the determinant of the connectivity matrix $\boldsymbol{W}$ of the linearized network is positive also implies $\det(\boldsymbol{J}) > 0$. Theoretical work on the stabilized supralinear network (SSN) has shown that the condition $\det(\boldsymbol{J}) > 0$ guarantees stable network dynamics assuming sufficiently fast inhibition (37), and plays an important role in determining aspects of neural dynamics such as bistability, persistent activity, and global oscillations (38).

Sadeh and Clopath (24) studied the conditions to obtain a suppressive, opposite-favoring mean perturbation response, and also concluded that disynaptic E → I → E connections must be sufficiently strong and like-to-like. Our results extend theirs in several ways. First, we are able to describe the spatial dependence of the response and not only the mean response. Second, we are able to obtain stronger and more precise mathematical results (for example, we do not require the assumption that inhibitory connections are much stronger than E → E connections, and in our analysis of feature-tuning, we considered the possibility of like-to-unlike I → I connections) by including stability constraints in our analysis. Finally, we note that the single-cell perturbation response data we and they are modeling (1) excludes neurons within 25 µm of the perturbed cell in lateral distance. Since the most nearby neurons studied were strongly excited, this exclusion is likely to result in an artifactual decrease of mean perturbation response, and thus it is unclear to what extent analysis based on mean perturbation response is valid. By additionally considering the dependence of response on distance between the perturbed neuron and the measured neuron, we are able to conclude that disynaptic E → I → E connections are sufficiently strong and like-to-like independently of the mean perturbation response.

We inferred parameters of the mean connectivity (*i.e.*, ignoring stochasticity in the connectivity) from optogenetic perturbation responses based on an explicit expression we derived for response vs. distance and feature preference, given that connectivity. Our approach is distinct from the works of (39–45), who inferred individual synaptic connections from whole-cell recordings of postsynaptic currents in response to perturbations of specific cells, based on models of monosynaptic intracellular responses. Other efforts to infer mean, and in some cases variance, of connectivity from responses to visual stimuli (16, 18, 28, 46–48) were either based on fitting by extensive search, or by comparison to expressions for responses that ignored space and/or feature dependence.

There are a few important future directions for our work. First, so far our analysis of the perturbation response equation has been restricted to a network with only a single inhibitory cell type and whose connectivity width depends only on presynaptic cell type. Without either restriction, our linear response equation would be composed of a sum of more than two spatial terms, which would make it difficult, if not impossible, to precisely characterize the conditions for the perturbation response to exhibit zero or more crossings in space. Thus, it remains to be seen whether analytic insight can be obtained into the behavior of more realistic models without these restrictions. Second, we have only considered models with dependence on a single feature. Mathematically it is straightforward to generalize our steady state solution to include an arbitrary number of periodic feature dependencies, but it is unclear how non-periodic features such as spatial and temporal frequency can be incorporated. Third, the feature tuning in our connectivity is parametrized by a cosine function, which fixes the feature tuning width. It will be important to investigate whether the theory can be adapted for other choices of feature tuning kernel that allow for variable feature tuning width, such as the wrapped Gaussian function. Finally, we have so far only dealt with single-cell perturbations in a linear network. Linearity is a reasonable approximation since a moderate single-cell perturbation is unlikely to generate significant nonlinear effects. However, many optogenetic experiments perturb an ensemble of neurons (2–5), or use one-photon methods to perturb large numbers of neurons and/or consider the combination of sensory and optogenetic stimuli (*e.g.*, 19, 28, 49), in which case nonlinear effects cannot be ignored. Furthermore, in nonlinear networks one also needs to consider the effects of connectivity disorder, which would both modify the mean perturbation response and potentially result in chaotic dynamics (19, 50). Thus, it is important to extend our work to consider nonlinear contributions to perturbation response.

Materials and Methods

Mathematical notation. Throughout the paper, scalar variables represented by lowercase letters like r, k. Vectors are represented by boldface lowercase letters such as $\boldsymbol{x}, \boldsymbol{k}$. Matrices are represented by boldface uppercase letters such as $\boldsymbol{\Sigma}, \boldsymbol{W}$. Given a matrix $\boldsymbol{W}$, its elements are written as W_{ij} or $[\boldsymbol{W}]_{ij}$, where the first notation is preferred whenever possible. Linear operators on vector spaces except $\mathbb{R}^n$ are represented by uppercase letters such as W, L, T.

Using the standard bra-ket notation, $|v\rangle$ represents a vector in a Hilbert space with label v. $\langle v|$ is the linear functional in the dual space associated with $|v\rangle$ such that $\langle v|(|u\rangle) = (|v\rangle, |u\rangle)$, where $(\cdot,\cdot)$ is the inner product on the Hilbert space. We write $\langle v \mid u\rangle$ to denote $\langle v|(|u\rangle)$. Similarly, given an operator T, we write $\langle v|T|u\rangle$ to denote $\langle v|(T|u\rangle)$. Given vectors $|v\rangle, |u\rangle$ in vector spaces V, U respectively, the vector $|v, u\rangle$ represents the vector $|v\rangle \otimes |u\rangle$ in the tensor product space $V \otimes U$.

Given cell type index $\alpha \in \mathbb{Z}_{N_c}$, we write $|\alpha\rangle$ to represent the standard basis vector $\boldsymbol{e}_\alpha \in \mathbb{R}^{N_c}$. Given vector $\boldsymbol{y} \in \mathbb{R}^d$, we write $|\boldsymbol{y}\rangle$ to represent the Dirac delta 'function' $\delta(\boldsymbol{x}-\boldsymbol{y})$. Similarly, given $\phi \in \mathbb{S}^1$, we write $|\phi\rangle$ to represent the Dirac delta 'function' $\delta(\theta-\phi)$ on the circle.

Model setup details. External input to the model (single-cell optogenetic perturbations) is modeled as a Dirac delta function. Specifically, external input due to the perturbation of neuron $(\beta, \nu, \boldsymbol{y}, \phi)$ is given by the equation

$$h_\alpha(\mu, \boldsymbol{x}, \theta) = hP_\alpha(\mu)^{-1}\delta_{\alpha\beta}\delta(\mu-\nu)\delta(\boldsymbol{x}-\boldsymbol{y})\delta(\theta-\phi) \quad [10]$$

where h is a scalar representing the perturbation strength, and the prefactor $P_\alpha(\mu)^{-1}$ ensures that the total input to the network $\int_0^1 \int_{\mathbb{R}^d} \int_{-\pi}^{\pi} h_\alpha(\mu, \boldsymbol{x}, \theta) P_\alpha(\mu)\, d\theta d\boldsymbol{x} d\mu$ is independent of the feature selectivity of the perturbed neuron ν.

We assume that the synaptic timescale of each neuron is only dependent on its cell type. Thus, the dynamical equation of our model is given by

$$(1+\tau_\alpha\partial_t) r_\alpha(\mu, \boldsymbol{x}, \theta, t) = \sum_{\beta=0}^{N_c-1} \int_0^1 \int_{\mathbb{R}^d} \int_{-\pi}^{\pi} W_{\alpha\beta}(\mu, \nu, \boldsymbol{x}-\boldsymbol{y}, \theta-\phi) \\ r_\beta(\nu, \boldsymbol{y}, \phi, t) P_\beta(\nu)\, d\phi d\boldsymbol{y} d\nu + h_\alpha(\mu, \boldsymbol{x}, \theta) \quad [11]$$

where τ_α is the time constant for cell type α. Stability of the network dynamics in general depends on the specific time constants chosen for each cell type. To simplify our discussion, however, we assume that the time constant of inhibitory neurons is sufficiently fast, such that the stability of the network dynamics depends only on the connectivity parameters (SI section 5C).

We define the linear response function, $\tilde{L}_{\alpha\beta}(\mu, \nu, \boldsymbol{x}-\boldsymbol{y}, \theta-\phi)$, as the solution $r_\alpha(\mu, \boldsymbol{x}, \theta)$ of the steady state equation 1 with external input $h_\alpha(\mu, \boldsymbol{x}, \theta)$ given by equation 10 where the scalar parameter h is set to 1 and $(\alpha, \boldsymbol{x}, \theta, \mu) \neq (\beta, \boldsymbol{y}, \phi, \nu)$. In terms of the linear operator $L = (I-W)^{-1}$, it can be written as

$$\tilde{L}_{\alpha\beta}(\mu, \nu, \boldsymbol{x}-\boldsymbol{y}, \theta-\phi) = P_\beta(\nu)^{-1}\langle \alpha, \mu, \boldsymbol{x}, \theta|L-I|\beta, \nu, \boldsymbol{y}, \phi\rangle \quad [12]$$

where the factor of $P_\beta(\nu)^{-1}$ comes from equation 10. The identity operator can be subtracted from L since we specified that $(\alpha, \boldsymbol{x}, \theta, \mu) \neq (\beta, \boldsymbol{y}, \phi, \nu)$.

Perturbation response in the full model. In equation 9 we specified the functional form of the perturbation response in the full model.

It contains a distance-dependent term, $\tilde{L}_{n\alpha\beta}(r)$, which is given by

$$\tilde{L}_{n\alpha\beta}(r) = \sum_{\rho=0}^{N_c^2-1} [\tilde{\boldsymbol{U}}_n \boldsymbol{P}_n]_{\alpha\rho} [\boldsymbol{P}_n^{-1} \tilde{\boldsymbol{V}}_n]_{\rho\beta} G_d(r; \lambda_{n\rho}) \qquad [13]$$

where $\tilde{\boldsymbol{U}}_n \in \mathbb{R}^{N_c \times N_c^2}, \tilde{\boldsymbol{V}}_n \in \mathbb{R}^{N_c^2 \times N_c}$ are matrices defined by

$$\tilde{U}_{n\alpha\gamma} = \sum_{\beta=0}^{N_c-1} A_{n\alpha\beta} \delta_{N_c\alpha+\beta,\gamma}, \quad \tilde{V}_{n\gamma\beta} = \sum_{\alpha=0}^{N_c-1} \delta_{N_c\alpha+\beta,\gamma}$$

$$A_{0\alpha\beta} = w_{\alpha\beta}\sigma_{\alpha\beta}^{-2}, \quad A_{1\alpha\beta} = w_{\alpha\beta}\sigma_{\alpha\beta}^{-2}\kappa_{\alpha\beta}. \qquad [14]$$

$\lambda_{n\rho} \in \mathbb{C}$ and $\boldsymbol{P}_n \in \mathbb{C}^{N_c^2 \times N_c^2}$ are defined such that $\boldsymbol{P}_n \boldsymbol{\Lambda}_n \boldsymbol{P}_n^{-1}$ is a diagonalization of $\boldsymbol{\Sigma}^{-1} - \tilde{\boldsymbol{V}}_n \boldsymbol{K}_n \tilde{\boldsymbol{U}}_n$, where $\boldsymbol{\Lambda}_n$ is the diagonal matrix of $\lambda_{n\rho}$, $\boldsymbol{K}_n \in \mathbb{R}^{N_c \times N_c}$ is defined by

$$K_{0\alpha\beta} = \delta_{\alpha\beta}, \quad K_{1\alpha\beta} = \delta_{\alpha\beta} \int_0^1 f_\beta(\mu) g_\beta(\mu) P_\beta(\mu)\, d\mu, \qquad [15]$$

and $\boldsymbol{\Sigma} \in \mathbb{R}^{N_c^2 \times N_c^2}$ is defined by

$$\Sigma_{\gamma\gamma'} = \delta_{\gamma\gamma'} \sum_{\alpha,\beta=0}^{N_c-1} \sigma_{\alpha\beta}^2 \delta_{N_c\alpha+\beta,\gamma}. \qquad [16]$$

These definitions arise naturally from the derivation of the perturbation response for the full model in SI section 2.

Equation 13 is completely analogous to the perturbation response of the simplified model given by equation 7, but it contains a sum over N_c^2 rather than N_c terms due to the fact that we now allow connectivity width to depend on both pre- and post-synaptic cell type rather than on pre-synaptic cell type alone.

Fitting of the spatial connectivity kernel. For Figure 1B, we combined the connection probability data from (11) and the connection strength data from (12) to estimate the product of connection probability and connection strength as a function of distance between excitatory and inhibitory neurons in mouse V1 L2/3. Since only I → E connection probability is measured in (11), we assumed that E → I connection probability is the same as I → E connection probability, an assumption supported by another dataset which shows that I → E and E → I connection probabilities have approximately the same width as a function of distance (20). Binning of connection probability and connection strength data is performed with bin edges from 0 to 500 μm spaced 25 μm apart. Given that connection strength is only measured between neurons up to about 100 μm apart in the data from (12), we assume that the connection strength for all bins in which no data is available is equal to the connection strength in the last bin in which data is available. The spatial kernel being fitted to this product is given by equation 3 with $d = 2$. For a given pair of post- and pre-synaptic cell types α, β, there are two free parameters: $w_{\alpha\beta}$ and σ_β. These parameters are fitted using the `optimize.curve_fit` function in the `scipy` Python library (51), which performs non-linear least squares. σ_β is initialized at 100 μm and $w_{\alpha\beta}$ is initialized to match the 2-norm of the data vector. Uncertainty of the fitted parameters is obtained from the default output of the `curve_fit` function, which estimates the covariance of fitted parameters by a linear approximation. This results in the best-fit parameters $\sigma_{\mathrm{E}} = (150.2 \pm 11.3)$ μm and $\sigma_{\mathrm{I}} = (107.6 \pm 8.4)$ μm.

Estimation of $r_0, r_0^{\min}$ from data. The 95% confidence intervals for $\frac{r_0}{\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}}}$ and $\frac{r_0^{\min} - r_0}{\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}}}$ in Figure 2D and 2E are estimated via bootstrapping. We independently sample each data point of the single-cell perturbation response curve in (1, Figure 2G) from a Gaussian distribution with its mean and standard error to obtain a random sample of the single-cell perturbation response curve. For each sample curve, we compute r_0 by linearly interpolating between the first two consecutive data points which exhibit a sign change. However, this would introduce a slight bias towards a smaller r_0 since the sampled curve may exhibit multiple crossings around r_0 and we are taking the first crossing. To address this bias we filter out all sampled curves with more than one crossing within 100 μm. We compute $r_0^{\min}$ as the location of the minimum of the sampled curve. However, the large standard errors in the data at large distances creates spurious minima in the sampled curve and thus introduces a small bias towards larger $r_0^{\min}$. To address this we simply consider the minimum of the sampled curve within 300 μm. We repeat the above procedures to obtain 100,000 samples of r_0 and $r_0^{\min}$. Finally, we divide each sample of r_0 and $r_0^{\min}$ by an independent sample of $\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}}$ using the mean and uncertainty of σ_{E} and σ_{I} as estimated in the Methods subsection Fitting of the spatial connectivity kernel, and compute the 2.5 and 97.5 percentiles of those 100,000 samples. This yields $\frac{r_0}{\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}}} \in (0.443, 0.691)$, $\frac{r_0^{\min} - r_0}{\sqrt{\sigma_{\mathrm{E}}\sigma_{\mathrm{I}}}} \in (0.260, 0.530)$

Comparison of theory and simulations. The parameters for the model in Figure 1D are given by Table 1. Since feature tuning preference in Figure 1D specifically refers to orientation tuning preference which is a variable in $[-\frac{\pi}{2}, \frac{\pi}{2})$ rather than $[-\pi, \pi)$, the connectivity function equation 8 as well as the linear response equation 9 need to be modified by replacing the factor of 2π by π and replacing $\cos(\theta - \phi)$ by $\cos(2(\theta - \phi))$.

Table 1. Model parameters for Figure 1D

Parameter	Value	Parameter	Value
σ_{EE}	125 μm	κ_{EE}	0.5
σ_{EI}	90 μm	$\kappa_{\mathrm{EI}}, \kappa_{\mathrm{IE}}$	−0.25
σ_{IE}	85 μm	κ_{II}	0.25
σ_{II}	110 μm	$f_\alpha(\mu), g_\alpha(\mu)$	μ
w_{EE}	3	$P_\alpha(\mu)$	1
$w_{\mathrm{EI}}, w_{\mathrm{IE}}$	4	τ_{I}	$\frac{1}{2}\tau_{\mathrm{E}}$
w_{II}	5.25		

For numerical simulations, the model is discretized on a regular grid with $N_x = 100$ by $N_y = 100$ spatial locations on a 1 mm × 1 mm torus ($d = 2$), $N_\theta = 12$ feature tuning preferences, and $N_\mu = 7$ feature selectivities. Spatial distances between neurons are measured by toroidal distances. The discretized model connectivity is obtained by multiplying the connectivity function equation 8 by a factor of

$$\Delta V = \left(\frac{1}{N_\mu}\right)\left(\frac{1\,\mathrm{mm}^2}{N_x N_y}\right)\left(\frac{\pi}{N_\theta}\right). \qquad [17]$$

To deal with the divergence of the spatial connectivity kernel $G_d(r; \lambda)$ for $d \geq 2$ as $r \to 0$, we simply set the connectivity strength between neurons at the exact same spatial location to 0. We provide a justification for this procedure in SI section 14. In other words, the discretized connectivity matrix $\boldsymbol{W}^{\mathrm{dis}}$ is defined by

$$W_{ij}^{\mathrm{dis}} = \begin{cases} W_{\alpha_i\alpha_j}(\mu_i, \mu_j, \boldsymbol{x}_i - \boldsymbol{x}_j, \theta_i - \theta_j)\Delta V, & \boldsymbol{x}_i \neq \boldsymbol{x}_j \\ 0, & \text{otherwise} \end{cases} \qquad [18]$$

where $\alpha_i, \mu_i, \boldsymbol{x}_i, \theta_i$ are the cell type, selectivity, spatial location, and feature preference respectively of neuron $i \in \{1, \cdots, N\}$ in the model, and $N := N_c N_\mu N_x N_y N_\theta$. Note that despite the multiplication by ΔV, the resulting discretized connectivity matrix is unitless since the connectivity function equation 8 has unit $[\mathrm{length}]^{-d}$. Network dynamics given by the discretized version of equation 11 is numerically integrated with order-5 Dormand-Prince method using the `torchdiffeq` package until convergence to steady state (52, 53), which is numerically determined by the condition $\left|\frac{\mathrm{d}r_i}{\mathrm{d}t}\right| \leq 10^{-5}|r_i| + 10^{-6}$ being satisfied for all $i \in \{1, \cdots, N\}$, where r_i is the firing rate of neuron i.

The analytical solution given by equation 9 is also scaled by the factor ΔV. Specifically, our analytical solution for the response of neuron i to the perturbation of neuron k in the discretized model is computed as

$$r_i^{\mathrm{dis}} = \begin{cases} h\tilde{L}_{\alpha_i\alpha_k}(\mu_i, \mu_k, \boldsymbol{x}_i - \boldsymbol{x}_k, \theta_i - \theta_k)\Delta V, & \boldsymbol{x}_i \neq \boldsymbol{x}_k \\ h\delta_{ik}, & \text{otherwise} \end{cases}. \qquad [19]$$

Intuitively, the scaling of the perturbation response by ΔV arises from the fact that the total external input does not increase with the number of neurons. We discuss this in detail in SI section 14.

Estimation of κ_{EE} from data. For the models fitted to experimental data in Figure 6, we estimate the value of the parameter κ_{EE} based on the publicly available mouse V1 L2/3 connection probability data from (11, Figure 2H), and use it to constrain fitted model parameters. To do this, we perform non-linear least squares using the `optimize.curve_fit` function in the `scipy` Python library (51) to fit the parameters a, κ_{EE} of the curve $\theta \mapsto a(1 + 2\kappa_{\mathrm{EE}} \cos(2\theta))$ to the connection probability data. Each data point is weighted inversely proportional to its standard error. The parameter a is initialized as the mean connection probability, while κ_{EE} is initialized at 0. κ_{EE} is constrained to be between -0.5 and 0.5. Uncertainty of the parameters is taken from the output of `curve_fit`. This yield a best-fit value of $\kappa_{\mathrm{EE}} = 0.198 \pm 0.054$.

Model fitting to experimental data. Here we describe the model fitting procedure used in Figure 6. The model consists of $N = 24{,}000$ neurons on a $900\,\mu\mathrm{m} \times 900\,\mu\mathrm{m}$ plane ($d = 2$). The cell type, spatial location, and orientation tuning preference of each neuron is randomly assigned, such that each neuron has $p_{\mathrm{E}} = 0.85$ probability of being an excitatory neuron and $p_{\mathrm{I}} = 0.15$ chance of being an inhibitory neuron, while spatial locations and tuning preferences are uniformly distributed. To avoid numerical issues due to the divergence of the spatial kernel $G_d(r;\lambda)$ as $r \to 0$, we require that all pairwise distances between neurons be at least $3\,\mu\mathrm{m}$. This is achieved by resampling the spatial location of one of the neurons from each pair of neurons whose pairwise distance is less than $3\,\mu\mathrm{m}$, and repeating until the requirement is satisfied. Since there is no experimental data for the perturbation response as a function of tuning selectivity, tuning selectivity is omitted in this model by setting $P_\alpha(\mu) = \delta(\mu - 1)$, *i.e.* every neuron is perfectly tuned. The inhibitory time constant is chosen to be twice as fast as excitatory time constant, *i.e.* $\tau_{\mathrm{I}} = \frac{1}{2}\tau_{\mathrm{E}}$. A random excitatory neuron within a $680\,\mu\mathrm{m} \times 680\,\mu\mathrm{m}$ window centered at the origin is chosen for perturbation. During model fitting, the steady-state response of the network is computed using the analytical solution. Specifically the steady state response of neuron i to perturbation of neuron k is computed as

$$r_i^{\mathrm{dis}} = \begin{cases} h\tilde{L}_{\alpha_i\alpha_k}(1, 1, \boldsymbol{x}_i - \boldsymbol{x}_k, \theta_i - \theta_k)\Delta V_{\alpha_k}, & \boldsymbol{x}_i \neq \boldsymbol{x}_k \\ h\delta_{ik}, & \text{otherwise} \end{cases} \qquad [20]$$

where $\Delta V_\alpha = \frac{1\,\mathrm{mm}^2 \cdot \pi}{p_\alpha N}$, $\alpha_i, \boldsymbol{x}_i$ are the cell type and spatial location of neuron i respectively, and $\tilde{L}_{\alpha\beta}$ is the linear response equation equation 9 with the factor of 2π replaced by π and $\cos(\theta - \phi)$ replaced by $\cos(2(\theta - \phi))$.

Model parameters are simultaneously fitted to both experimental data curves in Figure 6C, D. Neurons beyond a $680\,\mu\mathrm{m} \times 680\,\mu\mathrm{m}$ window centered at the origin are excluded to mimic the field-of-view of the experiment as well as to minimize boundary effects. Following the data analysis procedure of (1), neurons within $25\,\mu\mathrm{m}$ of the perturbed neuron are also excluded, and the mean responses of neurons within bins with bin widths of $60\,\mu\mathrm{m}$ are taken for fitting to the distance curve, while the mean responses of neurons within bins with bin widths of 25° are taken for fitting to the tuning preference curve. Since there are more data points for the distance curve and the y-values of the distance curve are an order of magnitude larger than the y-values of the tuning preference curve, to ensure both curves are equally well-fitted, we compute a weighted root-mean-square loss where the data points on each curve are weighted inversely proportional to the number of data points as well as the variance of the corresponding curve.

There are 13 relevant parameters for fitting: four connectivity strength parameters $w_{\alpha\beta}$, four connectivity width parameters $\sigma_{\alpha\beta}$, four feature tuning parameters $\kappa_{\alpha\beta}$, and the perturbation strength h. Since the perturbation strength h does not affect the shape of the response curve (response as a function of distance) and only affects its overall amplitude, we eliminate this parameter by normalizing both the model perturbation response as well as the data to unit norm during fitting. This leaves 12 free parameters $w_{\alpha\beta}, \sigma_{\alpha\beta}, \kappa_{\alpha\beta}$ which are fitted to minimize the loss. We impose several constraints on the 12 parameters during optimization. Specifically we constraint: 1) the signs of $w_{\alpha\beta}$ ($w_{\alpha\mathrm{E}} > 0$, $w_{\alpha\mathrm{I}} < 0$), 2) the magnitudes of $w_{\alpha\beta}$ to prevent unrealistically strong connections ($|w_{\alpha\beta}| < 10$), 3) the magnitudes of $\kappa_{\alpha\beta}$ to ensure compliance with Dale's law ($|\kappa| < 0.5$) 4) σ_{IE} and σ_{EI} to be within 2 standard deviations of the estimated values of σ_{E} and σ_{I} respectively as obtained from the Methods subsection Fitting of the spatial connectivity kernel, 4) σ_{EE} and σ_{II} to be between $75\,\mu\mathrm{m}$ and $175\,\mu\mathrm{m}$, 5) $\min\{\sigma_{\mathrm{EE}}, \sigma_{\mathrm{II}}\} > \max\{\sigma_{\mathrm{EI}}, \sigma_{\mathrm{IE}}\}$, based on connection probability data (20), 6) κ_{EE} to be within 2 standard deviations of the estimate value from the Methods subsection Estimation of κ_{EE} from data, 7) the network being an ISN ($w_{\mathrm{EE}} > 1$), and 8) the stability of the network dynamics (see SI section 5 on how the stability condition is approximately computed). Optimization is performed using the `optimize.minimize` function in the `scipy` library (51) with the SLSQP (Sequential Least SQuares Programming) method, with the gradient vector being computed with PyTorch's automatic differentiation engine (54).

Once the optimization algorithm has converged, the validation loss is computed as the weighted root-mean-square error (using the same weights as previously described) between the data and the average single-cell perturbation response obtained with 50 numerical simulations (5 random single-cell perturbations in 10 random instantiations of the model). This validation loss is further normalized such that a value of 1 is achieved by a model predicting zero perturbation response for every neuron. A random instantiation of the model is defined as a random assignment of the cell type and spatial location of each neuron, with connectivity strength from neuron j to neuron i defined by

$$W_{ij}^{\mathrm{dis}} = \begin{cases} W_{\alpha_i\alpha_j}(1, 1, \boldsymbol{x}_i - \boldsymbol{x}_j, \theta_i - \theta_j)\Delta V_{\alpha_j}, & \boldsymbol{x}_i \neq \boldsymbol{x}_j \\ 0, & \text{otherwise} \end{cases} \qquad [21]$$

where $W_{\alpha\beta}$ is the connectivity function equation 8 with 2π replaced by π and $\cos(\theta - \phi)$ replaced by $\cos(2(\theta - \phi))$. Each numerical simulation is performed using the same procedure as described in the Methods subsection Comparison of theory and simulations. If the network dynamics fail to converge for any one of the 50 simulations, the fitted parameters are discarded. This may occur despite the stability constraint imposed during optimization since the randomness of each neuron's spatial location causes variance in the spectral abscissa of the Jacobian that cannot be accounted for by our analysis of the continuum model. We also discard the fitted parameters if the validation loss is greater than 0.75.

To generate a reasonable distribution of fitted model parameters, instead of fitting the parameters directly to the mean perturbation response curve, we fit the parameters to a randomly sampled curve defined by the collection of points $\{(x_i, y_i)\}_i$, where y_i is an independent sample from the Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i)$ and μ_i, σ_i are the mean and standard error of the perturbation response at distance x_i respectively. Due to the large bin widths used in the data analysis procedure by (1), nearby data points on the perturbation response curves are correlated. This is addressed by simply only fitting the model to data points which are separated roughly $60\,\mu\mathrm{m}$ apart for the distance curve and 25° apart for the tuning preference curve. The optimization procedure is repeated with different random samples of the perturbation response curve, different random initializations of model parameters, and different random instantiations of cell types and spatial locations of neurons until 200 sets of fitted parameters are obtained. Since the optimization algorithm may sometimes be stuck at a local minimum of the loss function, only the top 50 models are kept.

Data, Materials, and Software Availability. Code for reproducing all figures is available at https://github.com/hchau630/chau-2024-exact.

ACKNOWLEDGMENTS. We thank Christopher Harvey and Selmaan Chettih for providing data from (1) for Figure 6C, D, and Petr Znamenskiy for providing data from (12) for Figure 1B. We thank Tuan Nguyen for providing feedback on the manuscript, and Mora Ogando for helpful discussions. This work is supported by the Gatsby Charitable Foundation (GAT3708; and GAT3850 to A.P), the Kavli Foundation, the NSF (DBI-1707398; and DGE-2036197 to H.Y.C.), the NIH (U01 NS108683, R01 EY029999, U19 NS107613; and T32 EY013933 to H.Y.C.), and the Simons Foundation (1156607, to A.P).

1. SN Chettih, CD Harvey, Single-neuron perturbations reveal feature-specific competition in V1. Nature **567**, 334–340 (2019).
2. IA Oldenburg, et al., The logic of recurrent circuits in the primary visual cortex. Nat. Neurosci. **27**, 137–147 (2024).
3. JH Marshel, et al., Cortical layer–specific critical dynamics triggering perception. Science **365**, eaaw5202 (2019).
4. L Carrillo-Reid, W Yang, Y Bando, DS Peterka, R Yuste, Imprinting and recalling cortical ensembles. Science **353**, 691–694 (2016).
5. L Carrillo-Reid, S Han, W Yang, A Akrouh, R Yuste, Controlling Visually Guided Behavior by Holographic Recalling of Cortical Ensembles. Cell **178**, 447–457.e5 (2019).
6. LE Russell, et al., The influence of cortical activity on perception depends on behavioral state and sensory context. Nat. Commun. **15**, 2456 (2024).
7. CE Deveau, et al., Recurrent cortical networks encode natural sensory statistics via sequence filtering (2024).
8. MV Tsodyks, WE Skaggs, TJ Sejnowski, BL McNaughton, Paradoxical Effects of External Modulation of Inhibitory Interneurons. The J. Neurosci. **17**, 4382–4388 (1997).
9. H Ozeki, IM Finn, ES Schaffer, KD Miller, D Ferster, Inhibitory Stabilization of the Cortical Network Underlies Visual Surround Suppression. Neuron **62**, 578–592 (2009).
10. A Sanzeni, et al., Inhibition stabilization is a widespread property of cortical networks. eLife **9**, e54875 (2020).
11. LF Rossi, KD Harris, M Carandini, Spatial connectivity matches direction selectivity in visual cortex. Nature **588**, 648–652 (2020).
12. P Znamenskiy, et al., Functional specificity of recurrent inhibition in visual cortex. Neuron **112**, S0896627323009728 (2024).
13. DB Rubin, SD Van Hooser, KD Miller, The Stabilized Supralinear Network: A Unifying Circuit Motif Underlying Multi-Input Integration in Sensory Cortex. Neuron **85**, 402–417 (2015).
14. R Rosenbaum, B Doiron, Balanced Networks of Spiking Neurons with Spatially Dependent Recurrent Connections. Phys. Rev. X **4**, 021039 (2014).
15. R Rosenbaum, MA Smith, A Kohn, JE Rubin, B Doiron, The spatial structure of correlated neuronal variability. Nat. Neurosci. **20**, 107–114 (2017).
16. M Dipoppa, et al., Vision and Locomotion Shape the Interactions between Neuron Types in Mouse Visual Cortex. Neuron **98**, 602–615.e8 (2018).
17. YN Billeh, et al., Systematic Integration of Structural and Functional Data into Multi-scale Models of Mouse Primary Visual Cortex. Neuron **106**, 388–403.e18 (2020).
18. S Di Santo, et al., The combination of feedforward and feedback processing accounts for contextual effects in visual cortex (2024).
19. A Sanzeni, et al., Mechanisms underlying reshuffling of visual responses by optogenetic stimulation in mice and monkeys. Neuron **111**, 4102–4115.e9 (2023).
20. L Campagnola, et al., Local connectivity and synaptic dynamics in mouse and human neocortex. Science **375**, eabj5861 (2022).
21. J Trousdale, Y Hu, E Shea-Brown, K Josić, Impact of Network Structure and Cellular Response on Spike Time Correlations. PLoS Comput. Biol. **8**, e1002408 (2012).
22. V Pernice, B Staude, S Cardanobile, S Rotter, How Structure Determines Correlations in Neuronal Networks. PLoS Comput. Biol. **7**, e1002059 (2011).
23. T Kanashiro, GK Ocker, MR Cohen, B Doiron, Attentional modulation of neuronal variability in circuit models of cortex. eLife **6**, e23978 (2017).
24. S Sadeh, C Clopath, Theory of neuronal perturbome in cortical networks. Proc. Natl. Acad. Sci. **117**, 26966–26976 (2020).
25. O Mackwood, LB Naumann, H Sprekeler, Learning excitatory-inhibitory neuronal assemblies in recurrent networks, (Neuroscience), Preprint (2020).
26. WW Hager, Updating the Inverse of a Matrix. SIAM Rev. **31**, 221–239 (1989).
27. KD Miller, A Palmigiano, Generalized paradoxical effects in excitatory/inhibitory networks (2020).
28. A Palmigiano, et al., Common rules underlying optogenetic and behavioral modulation of responses in multi-cell-type V1 circuits (2023).
29. H Ko, et al., Functional specificity of local synaptic connections in neocortical networks. Nature **473**, 87–91 (2011).
30. KD Miller, TW Troyer, Neural Noise Can Explain Expansive, Power-Law Nonlinearities in Neural Response Functions. J. Neurophysiol. **87**, 653–659 (2002).
31. D Hansel, C Van Vreeswijk, How Noise Contributes to Contrast Invariance of Orientation Tuning in Cat Visual Cortex. The J. Neurosci. **22**, 5118–5128 (2002).
32. NJ Priebe, D Ferster, Direction Selectivity of Excitation and Inhibition in Simple Cells of the Cat Primary Visual Cortex. Neuron **45**, 133–145 (2005).
33. PK LaFosse, et al., Cellular-resolution optogenetics reveals attenuation-by-suppression in visual cortical neurons. Proc. Natl. Acad. Sci. **121**, e2318837121 (2024).
34. E Persi, D Hansel, L Nowak, P Barone, C van Vreeswijk, Power-Law Input-Output Transfer Functions Explain the Contrast-Response and Tuning Properties of Neurons in Visual Cortex. PLoS Comput. Biol. **7**, e1001078 (2011).
35. Y Ahmadian, KD Miller, What is the dynamical regime of cerebral cortex? Neuron **109**, 3373–3391 (2021).
36. D Hansel, H Sompolinsky, Modeling Feature Selectivity in Local Cortical Circuits in Methods in Neuronal Modeling: From Ions to Networks. (MIT Press, Cambridge, Massachusetts), 2nd edition, (1999).
37. Y Ahmadian, DB Rubin, KD Miller, Analysis of the Stabilized Supralinear Network. Neural Comput. **25**, 1994–2037 (2013).
38. N Kraynyukova, T Tchumatchenko, Stabilized supralinear network can give rise to bistable, oscillatory, and persistent activity. Proc. Natl. Acad. Sci. **115**, 3464–3469 (2018).
39. AM Packer, et al., Two-photon optogenetics of dendritic spines and neural circuits. Nat. Methods **9**, 1202–1205 (2012).
40. CA Baker, YM Elyada, A Parra, MM Bolton, Cellular resolution circuit mapping with temporal-focused excitation of soma-targeted channelrhodopsin. eLife **5**, e14193 (2016).
41. OA Shemesh, et al., Temporally precise single-cell-resolution optogenetics. Nat. Neurosci. **20**, 1796–1806 (2017).
42. A Naka, et al., Complementary networks of cortical somatostatin interneurons enforce layer specific control. eLife **8**, e43696 (2019).
43. TA Hage, et al., Synaptic connectivity to L2/3 of primary visual cortex measured by two-photon optogenetic stimulation. eLife **11**, e71103 (2022).
44. Y Printz, et al., Determinants of functional synaptic connectivity among amygdala-projecting prefrontal cortical neurons in male mice. Nat. Commun. **14**, 1667 (2023).
45. MA Triplett, et al., Rapid learning of neural circuitry from holographic ensemble stimulation enabled by model-based compressed sensing (2022).
46. AJ Keller, et al., A Disinhibitory Circuit for Contextual Modulation in Primary Visual Cortex. Neuron **108**, 1181–1193.e8 (2020).
47. DP Mossing, J Veit, A Palmigiano, KD Miller, H Adesnik, Antagonistic inhibitory subnetworks control cooperation and competition across cortical space (2021).
48. N Kraynyukova, et al., In vivo extracellular recordings of thalamic and cortical visual responses reveal V1 connectivity rules. Proc. Natl. Acad. Sci. **119**, e2207032119 (2022).
49. JF O'Rawe, et al., Excitation creates a distributed pattern of cortical suppression due to varied recurrent input. Neuron **111**, 4086–4101.e5 (2023).
50. J Kadmon, H Sompolinsky, Transition to Chaos in Random Neuronal Networks. Phys. Rev. X **5**, 041030 (2015).
51. P Virtanen, et al., SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat. Methods **17**, 261–272 (2020).
52. J Dormand, P Prince, A family of embedded Runge-Kutta formulae. J. Comput. Appl. Math. **6**, 19–26 (1980).
53. RTQ Chen, Torchdiffeq (2018).
54. J Ansel, et al., PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation in Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. (ACM, La Jolla CA USA), pp. 929–947 (2024).